Academic Integrity in On-line Exams: Evidence from a Randomized Field Experiment

Flip Klijn
Mehdi Mdaghri Alaoui
Marc Vorsatz

This version: February 2021
(October 2020)

Barcelona GSE Working Paper Series

Working Paper nº 1210

Academic Integrity in On-line Exams: Evidence from a Randomized Field Experiment*

Flip Klijn† Mehdi Mdaghri Alaoui‡ Marc Vorsatz§

February 22, 2021

Abstract

We study academic integrity in a final exam of a compulsory course with almost 500 undergraduate students at a major Spanish university. Confinement and university closure due to Covid-19 took place by the end of the last lecture week. As a consequence, the usual classroom exam was turned into an unproctored on-line multiple-choice exam without backtracking. We exploit the different orders of exam problems and detailed data with timestamps to study students' academic integrity. Taking the average over questions that were part of both earlier and later "rounds," we find that the number of correct answers to questions in the later round was 7.7% higher than those to the same questions in the earlier round. Moreover, the average completion time of questions in the later round was 18.1% shorter than that of the same questions in the earlier round. We estimate that between 13.4% and 22.5% of the students cheated due to information flows from earlier to later rounds. Finally, a mere reminder of the university's code of ethics, which was sent to a subgroup halfway through the exam, did not affect cheating levels.

Keywords: field experiment, academic integrity, code of ethics, on-line exam, Covid-19.

JEL–Numbers: A22, C93, D9, I21, I23.

*We thank Rosemarie Nagel for useful comments and suggestions.

†Corresponding author. Institute for Economic Analysis (CSIC) and Barcelona Graduate School of Economics (Barcelona GSE), Campus UAB, 08193 Bellaterra (Barcelona), Spain; e-mail: flip.klijn@iae.csic.es. He gratefully acknowledges financial support from AGAUR–Generalitat de Catalunya (2017-SGR-1359) and the Spanish Ministry of Science and Innovation through grant ECO2017-88130-P AEI/FEDER, UE and the Severo Ochoa Programme for Centres of Excellence in R&D (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 PGC2018-096977-B-I00.

1 Introduction

Motivation and main results

In this paper, we present a randomized field experiment that aims to quantify the damage of cheating that is potentially caused by the absence of proctoring at on-line exams where students are only required to subscribe to the university's code of ethics. More specifically, we analyze data from a final exam at a major Spanish university in Spring 2020. The exam is an important part of a compulsory (but introductory) course on game theory for almost 500 undergraduate students (mostly in Economics and Business Management and Administration). The course took place in normal (classroom) circumstances until Covid-19 led Spanish authorities to decree confinement by the end of the last lecture week. As a consequence, the usual classroom exam was carried out using the university's on-line platform. This change opened the door to additional ways of potential cheating. The goal of our paper is to study academic integrity in this context by exploiting the specific design of the exam. In particular, in contrast to other studies, our analysis does not involve variables that are difficult to estimate or control for, e.g., latent ability.

In order to be able to summarize our main findings, we first describe the design of our exam. First, the exam consists of 20 multiple-choice questions which are grouped into five problems. Each (but the third) problem appears randomly at an earlier "round" (stage of the exam) for half of the students and at a later round for the other students. Second, since all students face exactly the same questions, the only difference between individual exams is the order of the questions and the order of the listed answers. Third, backtracking is not possible, i.e., once a student has moved to the next question, there is no possibility to go back to the previous question to change his/her answer.[1] We exploit the different orders of problems and detailed data with timestamps to study potential cheating by focusing on both correctness and completion times.

Our main findings are as follows. First, the students that received a given problem in the later round performed better than the other students in terms of both (higher) correctness and (shorter) completion time. Specifically, taking the average over questions that were part of earlier and later rounds, we find that the number of correct answers to questions in the later round was 7.7% higher than those to the same questions in the earlier round. And the average completion time of questions in a later round was 18.1% shorter than that of the same questions in an earlier round. Both comparisons are highly significant. We conjecture that mostly students from the later round profited from any answers and solutions that were shared through chat applications, social networks, or by e-mail/phone. Second, with respect to the questions of the problem that was not subject to order randomization, no significant differences regarding correctness and completion time are found for the different exam versions. Third, the reminder of the university's code of ethics, which was sent randomly to half of the students halfway through the exam, did not affect the correctness of the answers to nor the completion time of subsequent questions. We conjecture that most students just ignored the reminder as they believed that chances of being caught cheating were slim. We conclude that the university's code of ethics is not a very effective measure to reduce cheating.

[1] Here, "answer" includes the option of leaving the question unanswered.

While it is hard to provide an accurate estimate of the proportion of students that cheated at a particular question (or any subset of questions), we can estimate the additional proportion of students that cheated due to the different rounds of problems, which we refer to as additional cheaters/cheating. More precisely, additional cheating refers to cheating that originates from the flow of information from the earlier to the later round, i.e., on top of the potential communication during the same round. Assuming that cheaters answer correctly, a simple direct estimate of the proportion of additional cheaters is given by the proportion difference between later and earlier round students that provide a correct answer, which has an average of 0.056 and a maximum of 0.134.[2] A different estimate of additional cheating is obtained by additionally using completion times. Assuming that cheaters also do not purposely wait before answering a question they have obtained information about, we focus on students that give a correct answer in a very short period of time (meaning at least one standard deviation faster than the mean of the earlier round students). The average and maximum of the adjusted proportion difference between later and earlier students are given by 0.094 and 0.225, respectively, which can be considered upper bounds for the proportion of additional cheaters.[3] Our estimates thus suggest that between 13.4% and 22.5% of the later round students engaged in cheating at at least one question. In spite of (the evidence of) cheating, we find that exam grades are positively correlated with previous continuous assessment, and hence can be considered informative.

Related literature

Cheating and honesty have been studied in a wide range of contexts. Crawford and Sobel (1982) show theoretically that in games of strategic information transmission with self-interested players less information about the true state is revealed by the sender's message as the preferences between the players become less aligned. Kartik (2009) introduces fixed lying costs in the model of Crawford and Sobel (1982) in order to highlight that social interactions may not be reduced to a cost-benefit analysis over materialistic outcomes. A significant literature that uses laboratory experiments and that starts with Gneezy (2005) has corroborated this point of view by effectively showing that a non-negligible fraction of subjects is lying-averse when revealing their private information to others. Sánchez-Pagés and Vorsatz (2007) show experimentally that sanctions partially enhance honesty in this class of games. The related (but different) literature on cheating experiments introduced by Fischbacher and Föllmi-Heusi (2013) comes to a similar conclusion in the sense that a substantial proportion of subjects only partially misreport their private information to their advantage, maybe because of image concerns. In particular, Dufwenberg and Dufwenberg (2018) show theoretically that partial cheating can be sustained in equilibrium of the associated psychological game when players derive a disutility from being perceived to cheat. This literature thus suggests that many but far from all students would cheat in an exam setting when little to no consequences are to be expected.

Our study shares similarities with four recent field studies. First, Martinelli et al. (2018) analyze the extent of cheating in a governmental intervention program on incentivized learning

[2]The average and maximum are taken over all questions that are part of both an earlier and a later round.

[3]We refer to Section 3 for more details and a discussion of the second estimate.

in mathematics in Mexico. It is found that monetary incentives for students have a substantial effect on cheating: during the three year program, cheating is estimated to range between 5% and 7.5% in the absence and up to 32% in the presence of monetary incentives. Since cheating cannot be directly observed, the authors apply statistical methods from the education measurement literature[4] which, controlling for the ability of the students and the difficulty of the questions, aim at assessing for any ordered pair (i, j) of students from the same classroom, the probability that i copied from j. It is important to note that an advantage of our tailored design is that there is no need to specify a model of cheating behavior because additional cheating can be observed directly. Specifically, an essential feature of our designed experiment is that half of the students face problem I before problem II and the other half problem II before problem I, which allows us to control for latent variables such as ability. Thus, if ability caused that the group of students that solves problem I in the second round presents a higher correctness for problem I than the group of students that encounters problem I in the first round, then we should not find that the group of students that solves problem II in the second round presents a higher correctness for problem II than the group of students that gets problem II in the first round. However, since we do find that the correctness of the answers is higher for *all* problems when faced in later rounds, the data cannot be explained by ability alone.

Second, while Martinelli et al. (2018) concentrate on the correctness of the answers, Bilen and Matros (2020) consider correctness and completion time to provide evidence of cheating that took place in an on-line examination during a Covid-19 lockdown. More specifically, they analyze in detail the case of two students that present atypical time allocation to questions and extraordinary performance relative to midterm results. Based on their findings, the authors provide a policy recommendation that takes into account the issue of privacy concerns. Specifically, instructors should present students with two options: (1) If the student voluntarily agrees to use a camera to record themselves while taking an exam, this record can be used as evidence of innocence in the event that the student is accused of cheating, and (2) If the student refuses to use a camera due to privacy concerns, the instructor should be allowed to make the final decision on whether or not the student is guilty of cheating (with evidence of cheating remaining private to the instructor). Even though our experimental design allows us to estimate the magnitude of additional cheating, we recognize the difficulty, as pointed out by Bilen and Matros (2020), to fully detect whether particular students cheated. In particular, our estimation of the proportion of cheaters due to order effects should be interpreted in this light.

Third, Alan et al. (2020) present a designed field experiment on cheating in a creative performance task for 720 elementary school children. The authors distribute two different types of booklets with pre-drawn geometrical shapes, where the first type of booklets contains an ice cream and the second a microphone. The task of the children is to draw a meaningful figure in their booklet using circles and lines. Children sitting together on the same desk get different booklets and are explicitly told they should not look at each other's answers. In this setting, cheating occurs if a child with a pre-drawn ice cream (microphone) draws a microphone (ice cream). The authors have access to a wide range of individual and family characteristics and document that children

[4]For an overview, see, for example, Wollack and Fremer (2013) and Cizek and Wollack (2016).

with higher IQ and higher socioeconomic status have a higher likelihood of cheating. Materialistic incentives in the form of gift tokens, on the other hand, do not lead to more cheating. We do not find that the final exam grade is influenced by the individual characteristics (gender, risk aversion, attention levels) that were obtained via a voluntary (but weakly incentivized) questionnaire one week before the final exam.

Fourth, Vazquez et al. (2021) design a randomized field experiment to study the effects of proctoring on exam grades of two classes (face-to-face and on-line) of an introductory microeconomics course at a large university. Students whose exams were not proctored scored over 11% higher on average than those whose exams were proctored. However, the use of live proctors in the face-to-face class had a much larger effect on exam scores than web-based proctors in the on-line class. An important difference between Vazquez et al. (2021) and our study regards the treatment variables: we vary the order of problems and have an ethical reminder variable, while in Vazquez et al. (2021) students are either proctored or unproctored and the treatment groups are examined at different points in time.

Finally, a parallel education literature studies cheating and plagiarism within academia. For comprehensive overviews, we refer to the handbook McCabe et al. (2012) on cheating and to the handbook Bretag (2016) on breaches of academic integrity. Butler-Henderson and Crawford (2020) provide a systematic literature review of studies on on-line examinations, focusing on a variety of themes such as student perception, anxiety, student performance, and cheating. Many of these studies are based on surveys or interviews; and the studies that do perform a data analysis are not based on a specifically designed exam with a truly experimental setting such as ours.

Organization

The remainder of the paper is organized as follows. In Section 2, we describe the field experiment and our hypotheses. In Section 3, we present and discuss our results. Section 4 contains concluding remarks and policy recommendations. Appendices A, B, C, and D contain screenshots of the final exam, subject pool information, additional figures, and additional data analysis, respectively.

2 Randomized field experiment

Course structure and evaluation

The introductory course on game theory was distributed over 10 weeks in the second trimester of academic year 2019-2020. The final exam took place approximately 10 days later. The course is taught in English and is compulsory for all students in all four groups, which we denote by g1, g2, g3, and g4. Groups g1 and g2 are the groups of the bachelor's degree in Business Management and Administration as well as Economics. Group g3 corresponds to the double bachelor's degree in Law – Business Management and Administration/Economics. Finally, group g4 corresponds to the bachelor's degree in International Business Economics (whose program is fully taught in English). Each group had its own schedule of lectures (of this course and others). Students mostly attend the lectures of their own group and hence socialize mostly within their own group. One instructor was in charge of groups g1 and g2, and another instructor was in charge of groups g3 and g4.

Students' evaluation was based on continuous assessment. A student's final grade was determined by three items: 4 tests, 7 seminars, and 1 final exam. More specifically,

- each of the 4 tests could give up to 2.5 points [10 points in total],
- each of the 7 seminars gives 1 point for attendance, and up to 3 additional points for the participation/work in the 7 seminars [10 points in total], and
- the final exam could give up to 80 points.

The final grade (between 0 and 10) was obtained by dividing the number of achieved points by 10. The 4 tests took place through the university's on-line platform Moodle in weeks 3, 5, 7, and 9 of the course, and each of the tests consisted of 10 multiple-choice questions. Access to the test was open for approximately 24 hours, but once started with limited time to complete the test and without backtracking. The 7 (in-class) seminars took place in weeks $3, \ldots, 9$ 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 get up to 3 points in total). Students that missed 3 or more seminars were not allowed to take the final exam. Students were informed of all aforementioned details of the continuous assessment in the first week of the course. The decision to run an on-line final exam was only taken by the end of week 10 (when Spanish lockdown due to Covid-19 started), approximately 10 days before the scheduled final exam (day, time, and duration of the exam remained unchanged). The final exam was programmed and executed in Moodle.

A week before the final exam students were asked to fill out a short on-line survey (which took them on average 15 minutes). The survey consisted of a lottery task (Holt and Laury, 2002) to infer students' attitude towards risk and a 4-option multiple choice version of the cognitive reflection test (Frederick, 2005 and Sirota and Juanchich, 2018). Access to the survey was open during several days with the possibility of backtracking but limited time (30 minutes) once started. Students were informed that they could earn 2 additional points (on top of the established 100 points) by just filling out the survey but that the final total number of points for the course would be capped at 100. Approximately 98.2% of the students that participated in the final exam had completed the survey.

Design of final exam

The final exam consisted of 20 multiple-choice questions that were distributed over 5 problems. For each question we fixed five possible answers (of which only one was correct). For each question and for each student, the order of the five possible answers was chosen randomly. Selecting the correct answer gave 4 points, an incorrect answer 1 negative point, and not answering 0 points. Students did not receive any feedback until 2 weeks after the exam, when all grades were published and students could see the correct answers and check their answers and grade.

Next, we discuss the structure and relevant details of the final exam. Appendix A contains screenshots of the final exam. The first screen that the students saw was the part of the university's code of ethics that explicitly states:

> *"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, a student was provided with the exam instructions, which included information on the number of questions (20), the number of problems (5), a very brief generic description of each of the five problems, the number of points for a correct/incorrect/blank answer, and a reminder of the duration (120 minutes). Moreover, it was emphasized (in boldface) that *moving back to a previous question would not be possible*. Finally, students were informed that *after* the exam they would have the opportunity to participate in an experiment with a monetary prize, for which we added 3 more minutes to the duration of the on-line session.[5]

After clicking on the "continue" button at the bottom of the instructions, the first of the 20 multiple-choice questions appeared. Each subsequent question appeared on a new screen, but only after answering the previous question (or leaving it unanswered purposely). All students had the same 5 problems (and hence the same 20 questions), but problems and questions were permuted according to the scheme in Table 1. Students were not informed of the existence of different versions (permutations) of the exam. However, we believe that almost all students know that university rules establish that all students have to receive the same questions.

	problem in *r*ound						# students in *g*roup				
v	r1	r2	r3	reminder?	r4	r5	g1	g2	g3	g4	all
A	**I**	**II**	**III**	✓	**IV**	**V**	41	34	22	26	
	1,2,[3,4,5]	[6,7],[8,9,10],11	12,13		14,15,16	17,18,[19,20]					123
B	**I**	**II**	**III**	✗	**V**	**IV**	38	36	21	24	
	1,2,[3,4,5]	[6,7],[8,9,10],11	12,13		17,18,[19,20]	14,15,16					119
C	**II**	**I**	**III**	✗	**IV**	**V**	48	37	24	21	
	[6,7],[8,9,10],11	1,2,[3,4,5]	12,13		14,15,16	17,18,[19,20]					130
D	**II**	**I**	**III**	✓	**V**	**IV**	48	37	18	19	
	[6,7],[8,9,10],11	1,2,[3,4,5]	12,13		17,18,[19,20]	14,15,16					122
all							175	144	85	90	494

Table 1: Distribution of the 20 questions (labeled $1, \ldots, 20$) and the 5 problems (I, II, III, IV, V) in each of the 4 versions (A, B, C, D) and numbers of students. Questions within the same brackets were randomly permuted. "Reminder" refers to a reminder of the university's code of ethics. For instance, students that had version D of the exam received a reminder of the code of ethics right before they started working on problem V (which was their fourth problem, i.e., "round 4"), and "question 19" was their sixteenth or seventeenth question (depending on the individual draw by the on-line system).

[5]We implemented a variant of the die experiment introduced by Fischbacher and Föllmi-Heusi (2013). Specifically, students were randomly assigned to either of two experiments where each student was asked to roll one die or two dice using the website random.org. Students were asked to report the outcome (in the case of one die) or the sum of outcomes (in the case of two dice), say x. After the exam, and for each of the two experiments, one student was randomly chosen and received $2.5x$ (two dice) or $5x$ (one die) euros. Note that the die/dice experiment is independent of the exam, in terms of both grade and time. Moreover, the details of the experiment were only explained after the student had completed the exam. The design and findings are studied in a separate project.

Table 1 also contains information about the number of students in each of the four groups. Note that the random assignment of the four versions to the students led to an almost "balanced" distribution. More specifically, in each group the four versions of the exam were almost proportionally distributed over the students. In Appendix B, we present more detailed information about our subject pool. It is shown that the individual characteristics of the students (such as gender, risk aversion, attention levels, and performance in the previous continuous assessment) vary between groups. This is not surprising because the groups partly belong to different degrees and one should expect some self selection. More importantly though, there are no significant differences between the characteristics of the students that received different exam versions. Thus, randomization in the assignment of exam versions was successful and it is unlikely that our results are due to subject pool heterogeneity.

Hypotheses

Due to the very strict lockdown in Spain during the exam period it is very unlikely that students worked together on the exam in the same physical space. However, we could not impose any impediments to on-line communication (phone, email, social networks, etc.). We expect that (correct or incorrect) solutions/answers to any given question in the first round will accumulate and start to circulate so that students that are confronted with the same question in the second round are more likely to make a "more informed" decision, inducing more correct and/or quicker answers.

Formally, the *(average) correctness* of a given question in a particular round is defined as the proportion of correct answers to the question by the students that were faced with the question in that particular round. In this definition, leaving the question unanswered is considered an incorrect answer.[6] Similarly, the *(average) correctness* of a given problem in a particular round is defined as the average proportion of questions in the problem that are answered correctly by the students that were faced with the problem in that particular round. The *(average) completion time* of a question/problem in a particular round is the average time taken for the question/problem by the students that were faced with the question/problem in that particular round.[7] Finally, we will also study the correctness and completion time of a question/problem for a subpopulation of students in a particular round, in which case we indicate the subpopulation explicitly.

Hypothesis 1 (**Order effect: later round advantage**). *The answers to problems I and II depend on whether the problem appears in the first or second round. More specifically, for each of the problems I and II, the second round presents higher average correctness and shorter average completion time than the first round.*

One could naturally conceive a similar later round advantage for problems IV and V in rounds four and five. And, in fact, we will provide clear statistical evidence in this direction. However,

[6]The total number of times a question was left unanswered is 343, this amounts to $\frac{343}{20\times 494} = 3.47\%$ of the total number of answers.

[7]A question/problem is considered completed by a student if he/she moves to the next question/problem. So, completion time of a question can refer to the time used to read and think about a question but finally leaving it unanswered. The on-line platform measures time in minutes (where the minimum is 1 minute).

from an ex ante point of view, the results might be affected by the presence/absence of the ethical reminder, and we therefore decided to formulate Hypothesis 1 only with respect to the first two rounds.

In versions A and B, problem I was followed by problem II, while in versions C and D, problem II was followed by problem I. Since all students work on the same two problems in rounds 1 and 2 one can expect that a large group of students start working on problem III in (the common) round 3 around the same time.[8] Thus, the order of problems I and II should have no impact on the answers to problem III.

Hypothesis 2 (**Same preceding problems ⇒ similar answers in same new problem**). *There are no differences in the answers to problem III between versions A and B (history I,II) and versions C and D (history II,I): average correctness and average completion time are similar.*

At all exams of the university, students have to read (and sign) the university's code of ethics only once at the beginning (of the exam). Our reminder of the code of ethics halfway through the exam is exceptional. Thus, one could expect that the subsequent behavior of students that receive the reminder (versions A and D) is different from the students that do not receive the reminder (versions B and C). More specifically, one would expect that students that receive the reminder reduce possible engagement in communication with other students, which then is reflected in lower correctness and higher completion time of subsequent problems for this subpopulation.

Hypothesis 3 (**Disadvantage due to ethical reminder**). *Students that receive a reminder of the code of ethics [immediately after round 3] present lower average correctness and longer average completion time afterwards [in rounds 4 and 5] than the other students.*

3 Results

Analysis of hypotheses

It is convenient to first investigate Hypothesis 3 because the result we obtain will allow us to pool the data afterwards. As indicated in Table 1, after completing problem III in (common) round 3, the students with versions A and D received a reminder of the code of ethics, while the students with versions B and C did not. Table 2 provides aggregate data of the answers to the problems that were solved in subsequent rounds 4 and 5 (the last two rounds) and focuses on the effect of the reminder.[9] For instance, in the first row of Table 2, we compare the correctness and completion time of problem IV for students that are facing the problem in round 4 without a previous ethical reminder (version C) vs. with a previous ethical reminder (version A).

[8] If there is an order effect as hypothesized in Hypothesis 1, then the order (I, II) could be faster than (II, I) because problem II consists of 6 questions while problem I consists of only 5 questions. The reason is that solving a question requires more time than copying an answer. On the other hand, the difficulty of the questions could also shift the balance. So, it should be formally verified that the two groups of students indeed start working on problem III around the same time.

[9] Reported p-values are two-sided throughout the whole study.

problem	round	average correctness (proportion) no reminder	reminder	% *in*crease	p	average completion time (minutes) no reminder	reminder	% *de*crease	p
IV	r4	0.736	0.780	5.98	0.213	15.5	14.9	3.87	0.448
	r5	0.821	0.825	0.49	0.807	13.4	13.3	0.75	0.889
V	r4	0.884	0.891	0.79	0.809	21.3	21.4	-0.47	0.864
	r5	0.913	0.925	1.31	0.497	16.7	15.5	7.19	0.342

Table 2: Impact of no reminder vs. reminder (after round 3) of code of ethics on answers to problems IV and V (in rounds 4 and 5). The % increase/decrease is computed for "reminder" relative to "no reminder." We employ Mann-Whitney U tests at the student level.

It can be observed that the reminder of the code of ethics has two small effects that actually go in the direction *opposite* to that of Hypothesis 3: receiving a reminder of the code of ethics is associated with a (slightly) higher average correctness and (slightly) shorter average completion time.[10] Since there is no statistically significant comparison in Table 2 ($p > 0.213$ in all cases), we reject Hypothesis 3 and for the rest of our analysis we will pool observations independently of whether the student received (or not) a reminder of the code of ethics.

Result 1 (**No disadvantage due to ethical reminder**). *The average correctness and average completion time in rounds 4 and 5 of the students that receive a reminder of the code of ethics [immediately after round 3] are not statistically different from those of the students that do not receive a reminder of the code of ethics.*

Next, we investigate Hypothesis 1 on order effects. Since the reminder of the code of ethics turned out to be effectless, we consider the order effect not only for problems I and II (as stated in Hypothesis 1), but also for problems IV and V. Table 3 provides the aggregate data. In Table 3 and *throughout the paper we will use the following nomenclature*. Rounds 1 and 4 (rounds 2 and 5) 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 work on the problem in round 1, while the later group of students (or *later students*) refers to the group of students that work on the problem in round 2. Similarly, when discussing problem IV or V, the earlier/later group of students (or earlier/later students) refers to the group of students that work on the problem in round 4/round 5. Finally, in case of problem III, which is used as a control, "earlier" refers to versions A and B (history I, II) and "later" refers to versions C and D (history II, I).

problem	average correctness (proportion) earlier	later	% *in*crease	p	average completion time (minutes) earlier	later	% *de*crease	p
I	0.869	0.917	5.52	**0.000**	26.4	20.7	21.6	**0.000**
II	0.761	0.833	9.46	**0.000**	31.1	26.3	15.4	**0.000**
III	0.727	0.748	2.89	0.520	15.6	14.9	4.49	0.144
IV	0.758	0.823	8.58	**0.005**	15.2	13.3	12.5	**0.010**
V	0.888	0.919	3.49	**0.010**	21.3	16.1	24.4	**0.000**

Table 3: Impact of order of problems. The % increase/decrease is computed for "later" relative to "earlier." We employ Mann-Whitney U tests at the student level.

[10] Figures 6 (on problem IV) and 7 (on problem V) in Appendix C provide a visualization.

It is clear from Table 3 that for each of the two problems I and II, the students that are faced with the problem later achieve a higher average correctness and present a shorter average completion time than those that are faced with the problem earlier.[11] All results are statistically significant ($p < 0.001$ in the first two rows). Thus, we cannot reject Hypothesis 1.

Result 2 (**Order effect, problems I and II: later round advantage**). *The answers to problems I and II depend on whether the problem appears in the first or second round. More specifically, for each of the problems I and II, the second round presents a significantly higher average correctness and a significantly shorter average completion time than the first round.*

According to Table 3 the students who had problem I in round 1 and problem II in round 2 (versions A and B) required on average 52.7 minutes to complete the two problems, while the students who had problem II in round 1 and problem I in round 2 (versions C and D) required on average 51.8 minutes. In fact, we cannot reject the hypothesis that students facing versions A or B started problem III at the same time as the students with versions C or D ($p = 0.435$, Mann-Whitney U test). Thus, the third row in Table 3 shows that we cannot reject Hypothesis 2 on problem III: the small difference between the answers of students that had versions A and B (history I, II) and those that had versions C and D (history II, I) is not statistically significant at the 10-percent level. More specifically, the comparison between these two groups yields p-values of 0.520 and 0.144 for average correctness and average completion time, respectively.

Result 3 (**Same preceding problems ⇒ similar answers in same new problem**). *In terms of average correctness and average completion time, the answers to problem III of the students with versions A and B (history I, II) are not significantly different from the answers to problem III of the students with versions C and D (history II,I).*

Finally, since the reminder of the code of ethics turned out to be effectless and since we cannot reject that students with history I, II, III started problem IV at the same time as students with history II, I, III ($p = 0.106$, Mann-Whitney U test), it is possible to compare the answers to problems IV and V in the same way as problems I and II. We reach again the same conclusion: there is again a strong order effect in terms of average correctness, i.e., for each problem, the students that faced it in the last round have higher average correction and shorter average completion time ($p < 0.01$ in the last two rows of Table 3).

Result 4 (**Order effect, problems IV and V: later round advantage**). *The answers to problems IV and V depend on whether the problem appears in the fourth or fifth round. More specifically, for each of the problems IV and V, the fifth round presents a significantly higher average correctness and a significantly shorter average completion time than the fourth round.*

The scatter plot in Figure 1 complements Table 3 by providing a visualization of individual data.[12] In each panel/problem, each point[13] represents one student's proportion of correct answers

[11] Figures 8 (on correctness) and 9 (on completion time) in Appendix C complement Table 3.

[12] Figure 10 in Appendix C provides similar scatter plots for the four groups of students separately (see Table 1).

[13] A caveat is that students in the same group ("earlier" or "later") with the same number of correct answers and the same completion time (in minutes) are represented by overlapping circles or overlapping crosses.

and completion time. Specifically, the earlier students are represented by the circles ○, while the later students are depicted by the crosses +.

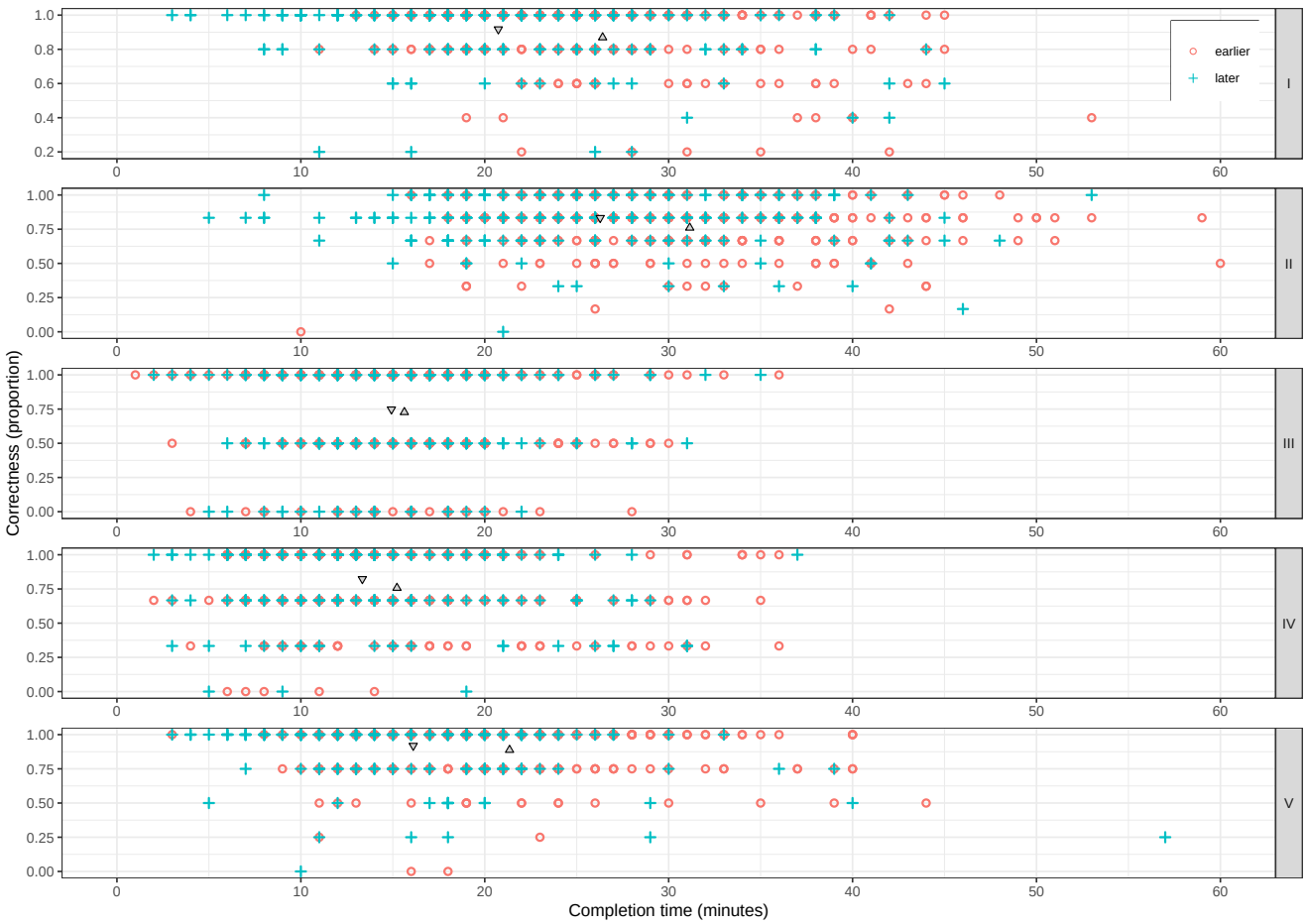

Figure 1: From top to bottom, panels describe correctness and completion time at the individual level for problems I, II, III, IV, and V separately. The overlap of at least one circle ○ and at least one cross + is visualized by a diamond-shaped form. The triangles △ and ▽ represent the averages of the earlier and later group, respectively.

For each of the problems I, II, IV, and V, we observe that in the top-left corner (i.e., higher correctness and shorter completion time) there are more crosses than circles, whereas in the bottom-right corner (i.e., lower correctness and longer completion time) there are more circles than crosses. As a consequence, the pair (average completion time, average correctness) of the later students is to the left of and above the corresponding pair of the earlier students, as indicated by the respective triangles △ and ▽ (which visualize part of the data in Table 3).

Analysis of order effect for individual questions

Next, we more closely study the order effect by checking whether it is driven by a limited number of particular questions. This analysis will also help us in estimating the proportion of students that engaged in cheating due to the different orders of the problems.

Table 4 provides aggregate data of the proportion of correct answers for each of the 20 questions separately. We observe that for almost all questions (17 out of 18) in problems I, II, IV, and V, the group of students that faces the problem later achieves a higher proportion of correct answers than the group of students that faces the problem earlier. The order effect is mostly apparent in questions where the earlier students achieve an average score that is below 0.8: questions 5, 7, 9, 10, 16, 20 (for each question, $p < 0.05$). The only exceptions are questions 8 and 15. Obviously, since scores are capped by 1, a high score in the earlier group does not allow for a much higher score in the later group. So, it is harder to find any effect of communication between the groups for these (most likely easy) questions. Also note that in line with Hypothesis 2 we do not observe statistically significant differences in either of the two questions (12 and 13) of problem III. Finally, by taking the average over questions that are part of earlier and later rounds (i.e., the questions in problems I, II, IV, and V), we find that the number of correct answers to questions in the later round is 7.7% higher than those to the same questions in the earlier round. The p-value of the binomial test that compares the correctness of the answers between earlier and later rounds (at the question level) is smaller than 0.0001.[14]

	problem I					problem II					
question	1	2	[3	4	5]	[6	7]	[8	9	10]	11
earlier	0.979	0.897	0.880	0.880	0.711	0.956	0.738	0.433	0.794	0.770	0.873
later	0.996	0.933	0.933	0.881	0.845	0.971	0.818	0.500	0.884	0.901	0.921
% increase	1.74	4.01	6.02	0.11	18.8	1.57	10.8	15.5	11.3	17.0	5.50
p	0.200	0.205	0.065	1.000	**0.000**	0.527	**0.042**	0.158	**0.009**	**0.000**	0.104

	problem III		problem IV			problem V			
question	12	13	14	15	16	17	18	[19	20]
earlier	0.702	0.752	0.949	0.715	0.609	0.946	0.967	0.867	0.772
later	0.718	0.778	0.963	0.776	0.730	0.964	0.960	0.881	0.870
% increase	2.28	3.46	1.48	8.53	19.9	1.90	-0.72	1.61	12.7
p	0.773	0.570	0.591	0.150	**0.006**	0.441	0.893	0.734	**0.006**

Table 4: Proportion of correct answers for each of the 20 questions separately: "earlier" vs. "later." The % increase is computed for "later" relative to "earlier." Questions between brackets [] were permuted at the individual level. We employ χ^2-tests of equal proportions.

Concerning the completion times, Table 5 provides aggregate data for each of the 20 questions separately.[15] We observe that for each question in problems I, II, IV, and V, the group of students that faces the problem later completes the question quicker than the group of students that faces the problem earlier. In fact, the order effect is significant at the 5-percent level for 15 out of the 18 questions. Comparing Tables 4 and 5, it follows that, except for question 7, any significant order effect in terms of the proportion of correct answers is accompanied by a significant order effect in terms of completion time. One possible explanation is that for any given question (especially for

[14]This result has to be taken with care because of an interdependent data structure: the answers to different questions come from the same students.

[15]Figure 11 in Appendix C complements Table 5.

an easy one) students in the later round can more easily reduce the time to answer it than increase the average proportion of correct answers (which might already be high in the earlier round). Note again that in line with Hypothesis 2 we do not observe statistically significant differences in either of the two questions (12 and 13) of problem III. Finally, regarding the questions in problems I, II, IV, and V, the average completion time of questions in the later round is 18.1% shorter than that of the same questions in the earlier round.[16] The p-value of the binomial test that compares the completion time of the answers between earlier and later rounds (at the question level) is smaller than 0.0001.[17]

	problem I					problem II					
question	1	2	[3	4	5]	[6	7]	[8	9	10]	11
earlier	3.77	4.81	4.68	7.08	6.07	3.76	4.61	4.83	4.26	7.40	6.27
later	3.04	3.67	4.01	5.47	4.55	3.62	4.55	4.29	3.43	5.48	4.90
% decrease	19.36	23.70	14.32	22.74	25.04	3.72	1.30	11.18	19.48	25.95	21.85
p	**0.000**	**0.000**	**0.014**	**0.000**	**0.000**	0.991	0.517	**0.002**	**0.001**	**0.000**	**0.000**

	problem III		problem IV			problem V			
question	12	13	14	15	16	17	18	[19	20]
earlier	7.15	8.48	2.87	6.15	6.21	5.35	3.66	5.57	6.77
later	6.93	8.00	2.51	5.57	5.27	4.04	2.49	4.31	5.27
% decrease	3.07	5.66	12.54	9.43	15.13	24.49	31.97	22.62	21.16
p	0.142	0.269	0.124	**0.049**	**0.023**	**0.000**	**0.000**	**0.001**	**0.000**

Table 5: Completion time for each of the 20 questions separately: "earlier" vs. "later." The % decrease is computed for "later" relative to "earlier." Questions between brackets [ ] were permuted at the individual level. We employ Mann-Whitney U tests.

Estimates of proportion of students involved in "additional cheating"

It seems hard to provide a reliable estimate of the proportion of students that cheated at a particular question (or any subset of questions). However, we can provide a proxy of the proportion of students that participated in *additional cheating*, which is the cheating due to the different orders of the problems. More precisely, additional cheating refers to cheating that originates from the flow of information from the earlier to the later round, i.e., on top of the potential communication during the same round.

A simple direct estimate of the proportion of students that participate in additional cheating can be obtained if we assume that cheaters answer correctly. Using the data on correctness in Table 4 of the questions in problems I, II, IV, and V we find that the average and maximum proportion difference of later vs. earlier students that provide a correct answer are 0.056 and 0.134, respectively.

[16] It can be hypothesized that the order effect that we observed for permuted problems (I and II, as well as IV and V) is also present for questions that were randomly permuted within a problem. We show in Appendix D that this order effect, which we call *instant order effect*, is almost negligible for correctness (the effect is significant in only 1 out of 44 instances) but more pronounced for completion time (the effect is significant in 13 out of 44 instances).

[17] Note again that the answers to different questions come from the same students.

A different estimate of additional cheating is obtained by additionally using completion times. To establish our second estimate, we make the additional assumption that cheaters do not purposely wait before answering a question they have obtained information about. This assumption seems reasonable as we have no indication that students were aware of the timestamps generated in the on-line platform: as far as we know, all final exams at the university were in-class until the week of our exam.[18]

Formally, let $i \in \{1, 2, \ldots, 20\} \backslash \{12, 13\}$, i.e., the questions that appear in two different (earlier and later) rounds. Let μ_{ie} and σ_{ie} be the mean and standard deviation of the completion time of question i for the earlier group of students. A student (from the earlier or later group) is considered to give a quick answer to question i if his/her completion time is in the interval $[0, \mu_{ie} - \sigma_{ie}]$. Our definition of "quick answer" only takes into account the earlier group of students because the problem appears first for (only) these students and hence their completion times are more realistic than those of the later group, in the sense that they are less affected by information flow.[19]

We do not exclude the possibility that some cheaters in the earlier group do not answer quickly, as they might have to wait until an answer and more information become available to them. However, once information starts to flow, it is likely that it reaches students from the later group as well, even before the latter students are facing the question in the later round. Hence, it is less likely that cheaters in the later group have to wait, if at all, before information about the question becomes available.

Students that answer correctly and quickly either cheat or do not cheat. The second category consists of "lucky gamblers" (no cheat, without knowledge) and students with high latent ability (no cheat, with knowledge). Then, since the different versions of the exam were randomly assigned to students, a higher proportion of correct and quick answers in the later group relative to the earlier group can only be explained by additional cheating. Formally, let p_{ie} (p_{il}) be the proportion of earlier (later) students that answer question i correctly and quickly. Then, the difference $p_{il} - p_{ie}$ in the two proportions is an estimate of additional cheating. However, it should be considered an upper bound because of the previously mentioned possibility that some cheaters in the earlier group do not answer quickly.

Figure 2 graphically depicts the proportions p_{ie} and p_{il} and indicates whether the difference is statistically significant. We observe that proportions are (almost always) larger for the later students and the difference is often significant. Importantly, no significant differences are found for questions of problem III.[20] By taking the average/maximum over the questions in problems I, II, IV, and V, we obtain an average and maximum proportion of 0.094 and 0.225, respectively,

[18] Over the last few years, only short intermediate tests, questionnaires, and surveys were carried out through the on-line platform. However, we are not aware of courses where instructors extracted timestamps from the platform, let alone shared this information with students. Moreover, the extraction of timestamps is non-trivial and requires substantial mechanical labor.

[19] Another possibility would be to define "quick answer" based on the completion time for the earlier students that give a correct answer. However, since the final estimate is virtually the same, we omit the corresponding analysis.

[20] We include questions 12 and 13 in Figure 2 as a control. Recall that questions 12 and 13 belong to problem III, which all students face around the same time. In these two cases "quick answer" is based on the students that had a particular history (I,II) of problems.

which constitute our second estimates (upper bounds) of additional cheating.

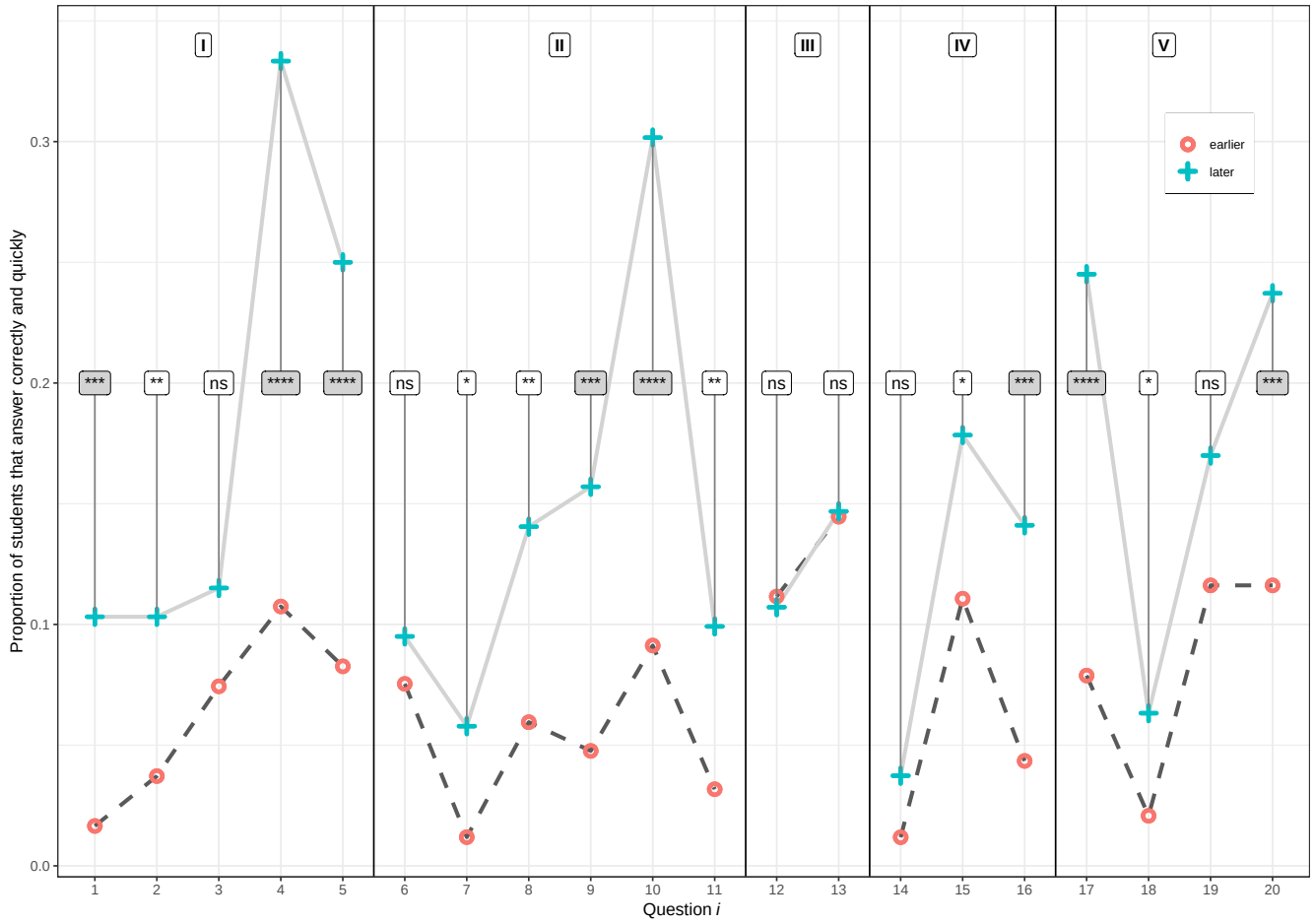

Figure 2: The proportion of students in the earlier (later) group that answer both correctly and quickly is indicated by ∘ (+). $^{*}\,p < .05$; $^{**}\,p < .01$; $^{***}\,p < .001$; $^{****}\,p < .0001$; ns=not significant. We employ χ^2-tests of equal proportions.

We summarize our insights regarding the proportion of additional cheaters in the next result.

Result 5 (**Proportion of students involved in additional cheating**). *Regarding the questions that appear in different (earlier and later) rounds, the average (maximum) proportion of students that engage in additional cheating is of the order 0.056-0.094 (0.134-0.225).*

If cheating at (at least) one question makes a student a "cheater," then Result 5 suggests that between 13.4% and 22.5% of the students can be labeled as "cheaters." Each of these two percentages are obtained with respect to some specific later round group (and question), but there is no reason to believe that the other students would behave differently if they would be in the same situation. So, it seems reasonable to interpret the two percentages as if they were percentages with respect to the whole population of examinees. Obviously, we are silent about cheating that takes place exclusively due to information flow in the same round. Therefore, the total percentage of students that engage in cheating is likely to be higher.

Analysis of informativeness of exam grades

In view of our evidence of substantial cheating at the exam, an important question is whether the final exam grades are still informative, i.e., positively related to the students' latent abilities. To answer this question, 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 4 intermediate tests). Note that while attendance/performance at the seminars is basically cheating-proof, this is not necessarily the case for the (on-line) intermediate tests. However, the weight of the tests was limited (10% of the final grade) and students received randomly one of the two versions of each question in the tests. Thus, it is reasonable to assume that the intensity/extent of cheating in the tests is smaller than in the final exam. In short, the continuous assessment can be considered to be an acceptable proxy of students' latent abilities. Table 6 summarizes the estimation results.

Intercept	26.5690^{***}
	(6.3128)
Intermediate tests	1.6246^{***}
	(0.3060)
Seminar attendance	3.2755^{***}
	(0.9281)
Seminar participation	0.7175
	(0.5765)
Gender	0.1332
	(1.0534)
Risk aversion	0.2922
	(0.4464)
Attention	0.7508^{*}
	(0.4518)
Observations	464

Table 6: Random effects estimation of the dependency of the final exam grade. We refer to Appendix B on subject pool information for a formal definition of the explanatory variables. The model includes group dummies ($g2$, $g3$, and $g4$) and dummies for the different exam versions (B, C, and D). In parenthesis, we present standard deviations. $^{*}\,p < .1$; $^{**}\,p < .05$; $^{***}\,p < .01$. If the dummies variables are removed from the model, participating actively in seminars becomes significant at the 1-percent level, while attention is not significant anymore.

We observe that there is a positive, highly significant relationship between the final exam grade and both the grade in the intermediate tests and the seminar attendance (all expressed in "points," as explained in "Design of final exam" in Section 2). Hence, final exam grades can be considered informative. Note that participating actively in the seminars is, as expected, also positively correlated with the final exam grade, but not significantly so. In the estimation, we control for three individual characteristics: gender, risk aversion (Holt-Laury lottery task), and

attention levels (cognitive reflection test). Among these, only the latter is significant at the 10-percent level.

4 Concluding remarks

Our study provides clear evidence of cheating behavior in an on-line final exam at a major Spanish university. Is it possible to reduce cheating in the setting of multiple-choice questions? We believe that in the case of *in-class* exams, the *fairest* procedure is to provide all students with the same questions: no student can complain that he/she failed or underperformed relative to peers because he/she had an "unlucky draw" of questions. However, giving all students the same questions (especially if there are no further measures to inhibit cheating) seems a risky procedure for on-line exams. In fact, a fair and possibly more cheating-proof procedure in this case would be precisely the opposite of a unique list of questions: for each question, a sufficiently large number of different versions should be generated and then randomly assigned to students. Here, "different versions" refers to scaling, switching, etc. of numerical values, and depending on the permitted procedures by the university's authorities, a potentially wider range of variations. Thus, if the number of questions is large enough, the random draws for each question will generate individual exams of a similar over-all level of difficulty. We leave for future research the potential relation between the number of different versions for each question and the mitigation of cheating.

We find that a reminder of the university's honor code does not reduce cheating. Using self-reported questionnaires, Gurung et al. (2012) assess how likely students are to cheat under eight different honor pledges. They find that honor pledges that are explicit about the consequences (such as academic hearings, suspensions, or expulsions) of breaches of the examination rules reduce students' a priori propensity to cheat. McCabe et al. (2002) find that honor codes that include practices such as unproctored exams, written pledges, and hearing bodies can reduce cheating. The discrepancy with respect to our findings is possibly explained by peer effects and enhanced communication methods. If students supposedly learn before our on-line exam that there will be cheating —for example, they might observe that communication channels are established in order to pass on information during the exam—, they might feel that cheating is justified as it avoids getting worse grades than other students.

The university's code of ethics that was used in our exam does not specify any such consequences. The university very recently renewed its code of ethics. However, the new text is still silent about the possible consequences of breaches of examination rules. We believe that this a missed opportunity even though the precise working of honor pledges in on-line exams is not yet fully understood. On the one hand, from a purely materialistic point of view, cheating should be expected if there are little to no consequences or if students are not aware of them. On the other hand, intrinsic motivations and social norms shape human behavior in many socioeconomic environments. More data from randomized experiments is needed to evaluate the effectiveness of honor pledges.

Finally, apart from directly analyzing the correctness of students' answers, we have employed completion times as a tool to search for anomalies in students' answers. As we mentioned in the

previous section, the students that participated in our final exam were most likely not aware of the timestamps generated in the on-line platform. Of course, when students learn about instructors verifying timestamps to detect cheaters, they might gradually opt for more sophisticated cheating behavior, for instance by including extra waiting time before answering (difficult) questions. Thus, completion times could become a less useful tool to analyze cheating.

A 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 on-line platform). The first screen that the students saw contains part of the university's code of ethics and is reproduced in Figure 3.

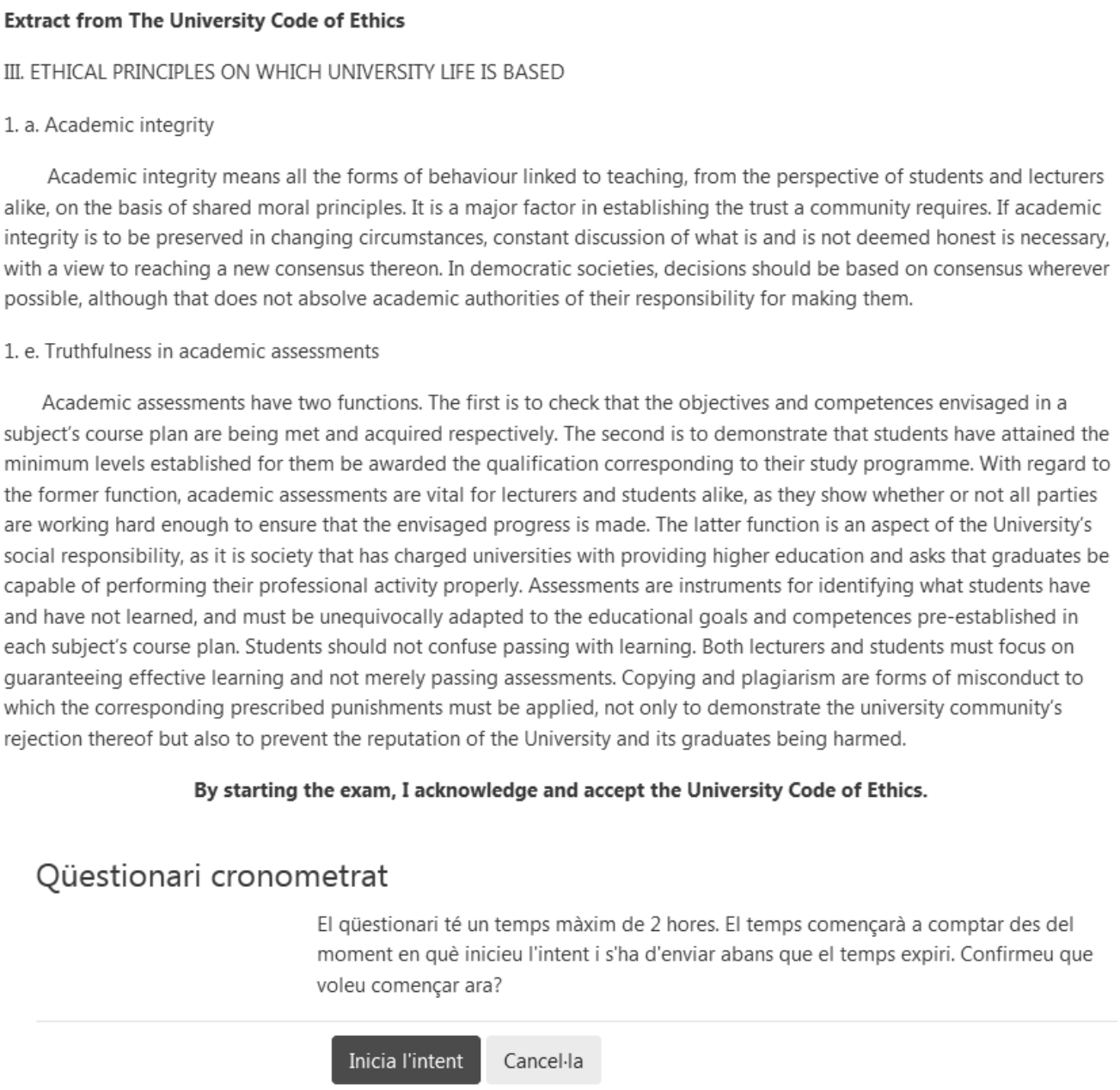

Figure 3: Part of the university's code of ethics.

The last part reads 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]"

After starting and subscribing to the code of ethics, a student was immediately provided with the exam instructions and further information, as shown in Figure 4.

IMPORTANT: read the instructions below before you proceed to question 1.

Beginning of instructions.

Since you have proceeded and started the exam, you have subscribed to the University Code of Ethics.
The exam consists of 20 multiple-choice questions.
The 20 multiple-choice questions are distributed over 5 problems.

- a problem on decision making with risky alternatives;
- a problem on a normal-form game;
- a problem on a zero-sum game;
- a problem on an extensive-form game;
- a problem on a market with two competing firms.

Some questions are easier or require less calculations than other questions. Do not waste unnecessary time on easy questions.
The unique correct answer to each question gives **4 positive points**. Each incorrect answer gives **1 negative point**. For instance, if Pep gives 15 correct answers, gives 3 incorrect answers, and keeps 2 questions unanswered, then he obtains $15\times4 + 3\times(-1) = 57$ points (out of 80 points).
You have one attempt and the official 120 minutes to submit all your answers. **You have to answer each question before you move to the next one. In other words, moving back to a previous question is not possible.**
After finishing the exam (i.e., the 20 questions), you will have the opportunity to participate in an experiment with a monetary prize. The experiment is presented as "question 21" but **does not count** for your grade (nor affect it in any possible way). If you participate, you can earn some money. (If you do not participate, you will not earn any money for sure). You will need less than 2 minutes to answer question 21. However, we decided to add 3 minutes to the 120 minutes of the exam so that you have sufficient time for everything.

End of instructions.

Figure 4: Exam instructions.

Students that had been (randomly) assigned versions A and D of the exam were reminded of the university's code of ethics after completing problem III, as shown in Figure 5.

Reminder.

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.

End of reminder.

Figure 5: Reminder of the university's code of ethics.

B Subject pool information

	Gender	Risk aversion	Attention	Intermediate	Attendance	Participation
Group 1	0.581	4.357	2.006	5.518	6.878	2.006
Group 2	0.451	4.340	1.962	5.196	6.777	2.748
Group 3	0.561	4.134	2.359	6.425	6.752	2.595
Group 4	0.666	4.226	1.906	7.349	6.920	2.986
Exam version A	0.555	4.085	2.025	5.998	6.777	2.461
Exam version B	0.558	4.342	2.018	5.665	6.900	2.531
Exam version C	0.566	4.350	2.183	6.054	6.791	2.516
Exam version D	0.534	4.379	1.948	5.844	6.862	2.465

Table 7: Subject pool information (average). The personal characteristics are defined as follows: Gender (1 for female), Risk aversion (number of risky choices in the Holt-Laury lottery task [0-10]), Attention (number of correct choices in the cognitive reflection test [0-3]), Intermediate (total number of points obtained in the 4 intermediate tests [0-10]), Attendance (total number of seminars attended [0-7]), Participation (average points obtained for participating in seminars [0-3]).

Since the groups partly correspond to different degrees, it is to be expected that the groups are not identical in their personal characteristics. For example, it turns out that the proportion of females in group 2 (0.451) is significantly smaller than that in group 4 (0.666). The two-sided p-value of the corresponding Mann-Whitney U test is 0.0028. For our field experiment it is important that the individual characteristics do not vary in the randomly assigned exam version. Among all possible pairwise comparisons (6 for each personal characteristic), the lowest two-sided p-value is 0.0888, which corresponds to the comparison of the number of correct answers in the cognitive reflection test between students with exam version A and students with exam version B. Since the two-sided p-value of all other pairwise comparisons is at least 0.1632, this is the only comparison that is significant at the 10-percent level.

C Additional figures

Figures 6 and 7 complement Table 2. They focus on the effect of the reminder of the code of ethics (after completing problem III in (common) round 3) on problems IV and V. It can be observed in the second column of the two figures that in terms of correctness the group of students without a reminder never first-order stochastically dominates the group of students with a reminder, i.e., the dark-gray graph is never completely situated below the light-gray graph. In fact, in case of the earlier group of problem IV, the group of students with a reminder first-order stochastically dominates the group of students without a reminder.[21] This provides further support for the rejection of Hypothesis 3.

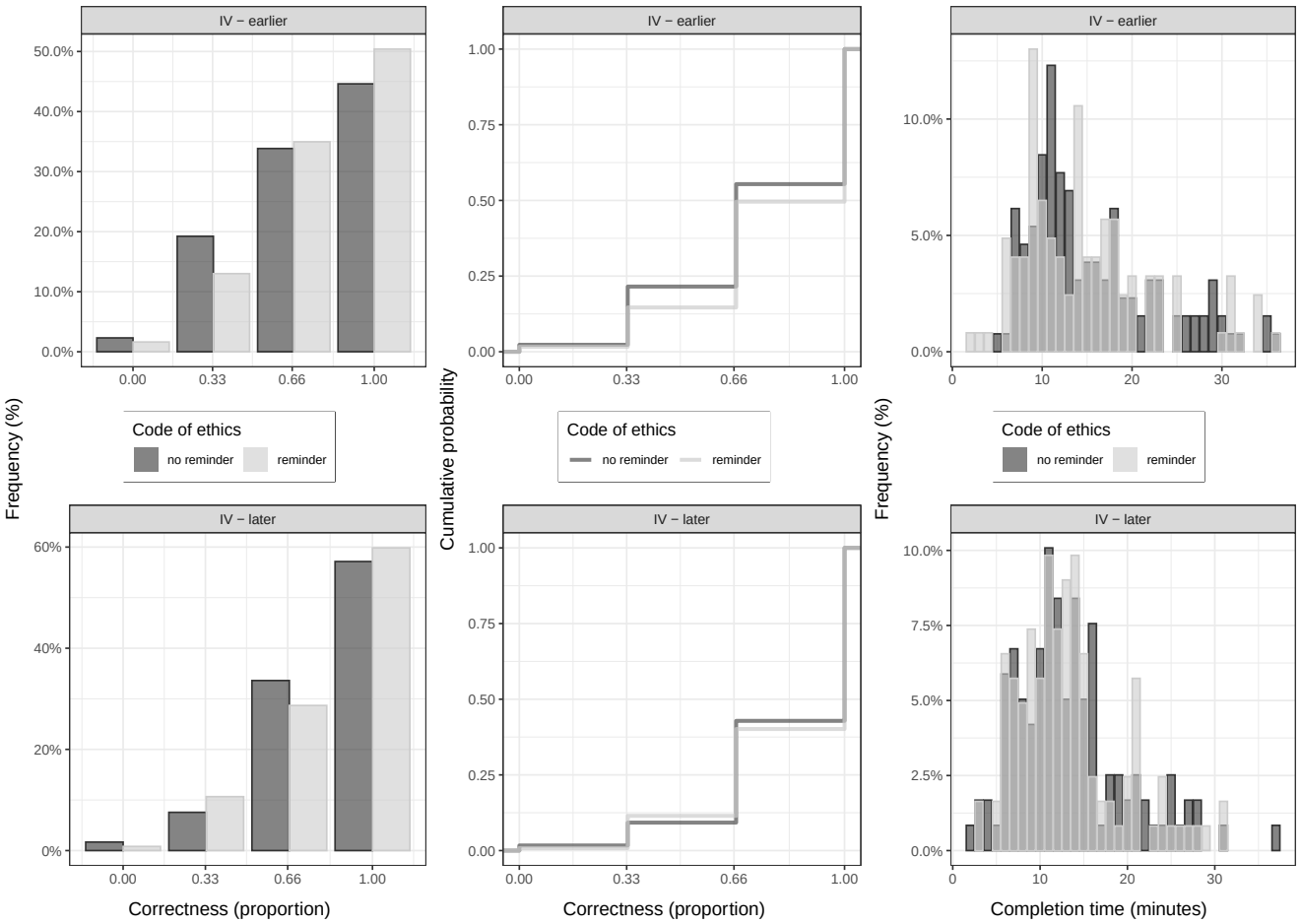

Figure 6: Problem IV. The first (second) row corresponds to the earlier (later) group. The first (second) column presents frequencies (cumulative probabilities) of the correctness. The third column presents frequencies of the completion time. The "intermediate-gray" in the third column represents "reminder" when "no reminder" has a higher frequency and it represents "no reminder" when "reminder" has a higher frequency.

[21] One can almost draw the same conclusion for the later group of problem V.

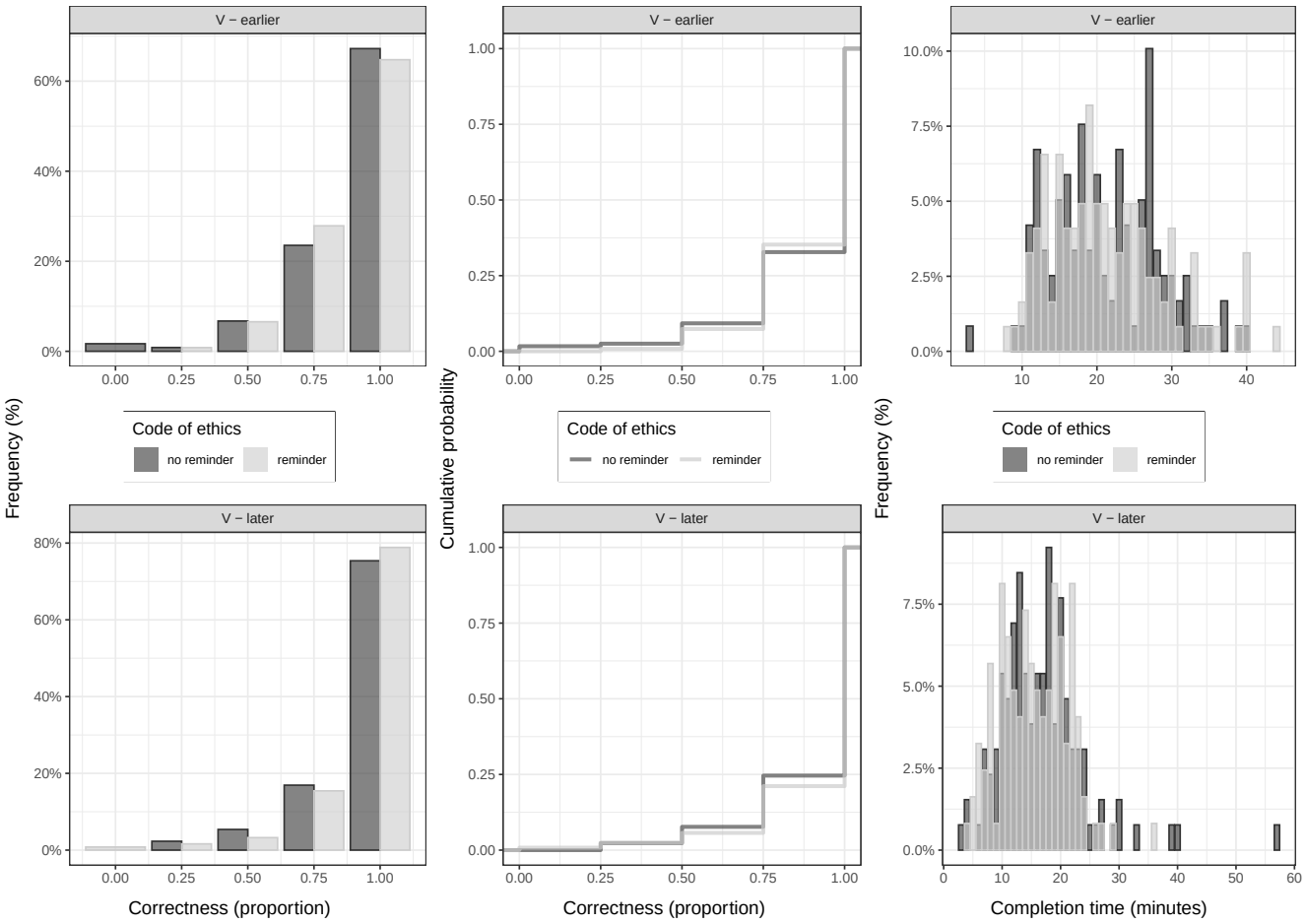

Figure 7: Problem V. The first (second) row corresponds to the earlier (later) group. The first (second) column presents frequencies (cumulative probabilities) of the correctness. The third column presents frequencies of the completion time. The "intermediate-gray" in the third column represents "reminder" when "no reminder" has a higher frequency and it represents "no reminder" when "reminder" has a higher frequency.

Figures 8 (on correctness) and 9 (on completion time) below complement Table 3. Concerning the right hand side of Figure 8, it can be observed that for problems I, II, IV, and (almost) V, the later group first-order stochastically dominates the earlier group in terms of correctness: the light-gray cumulative distribution function is always below the dark-gray cumulative distribution function. This constitutes further support for the order effect stated in Results 2 and 4.

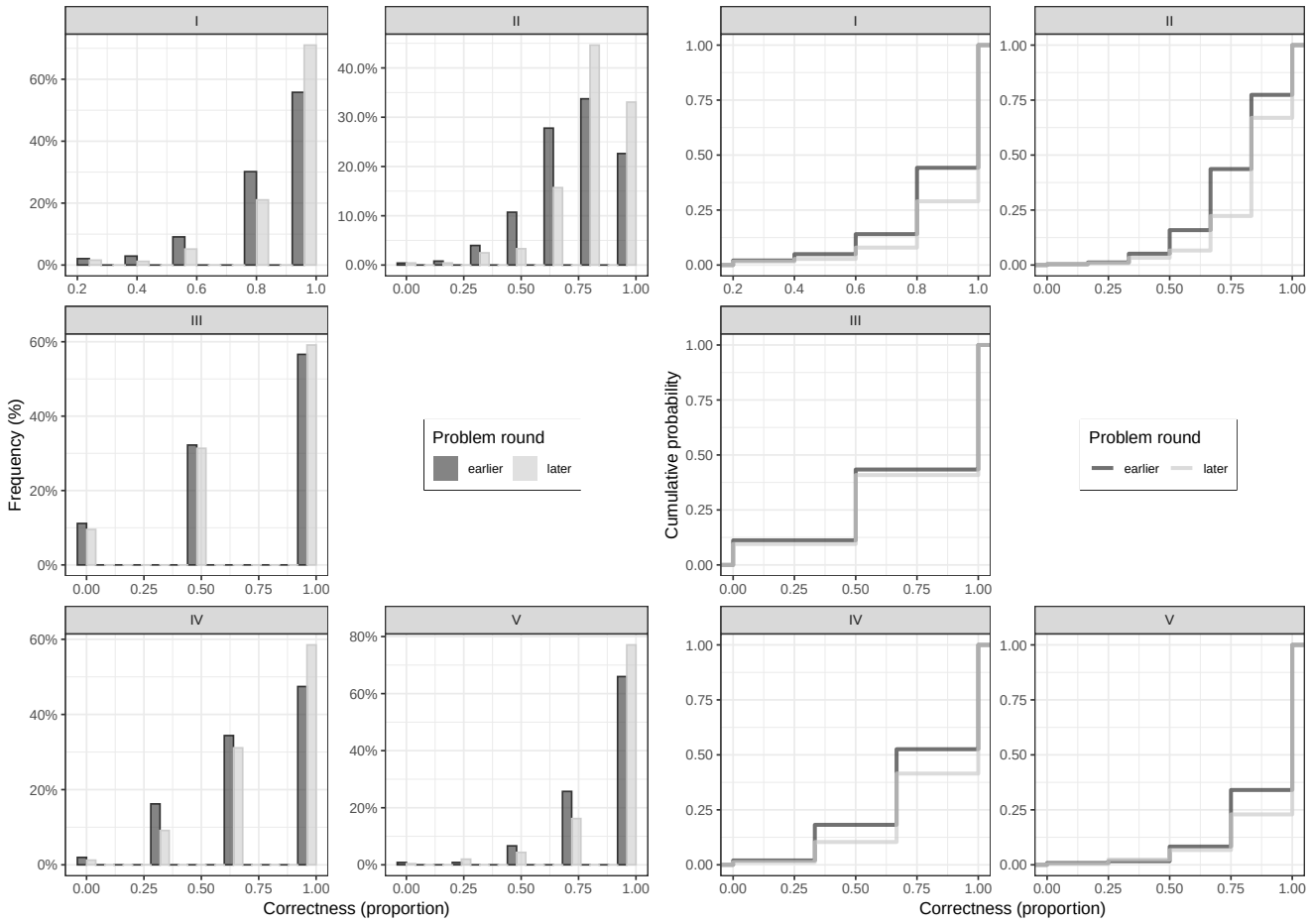

Figure 8: Correctness of each of the problems I, II, III, IV, and V. The left hand side presents frequencies while the right hand side depicts the corresponding cumulative distribution functions.

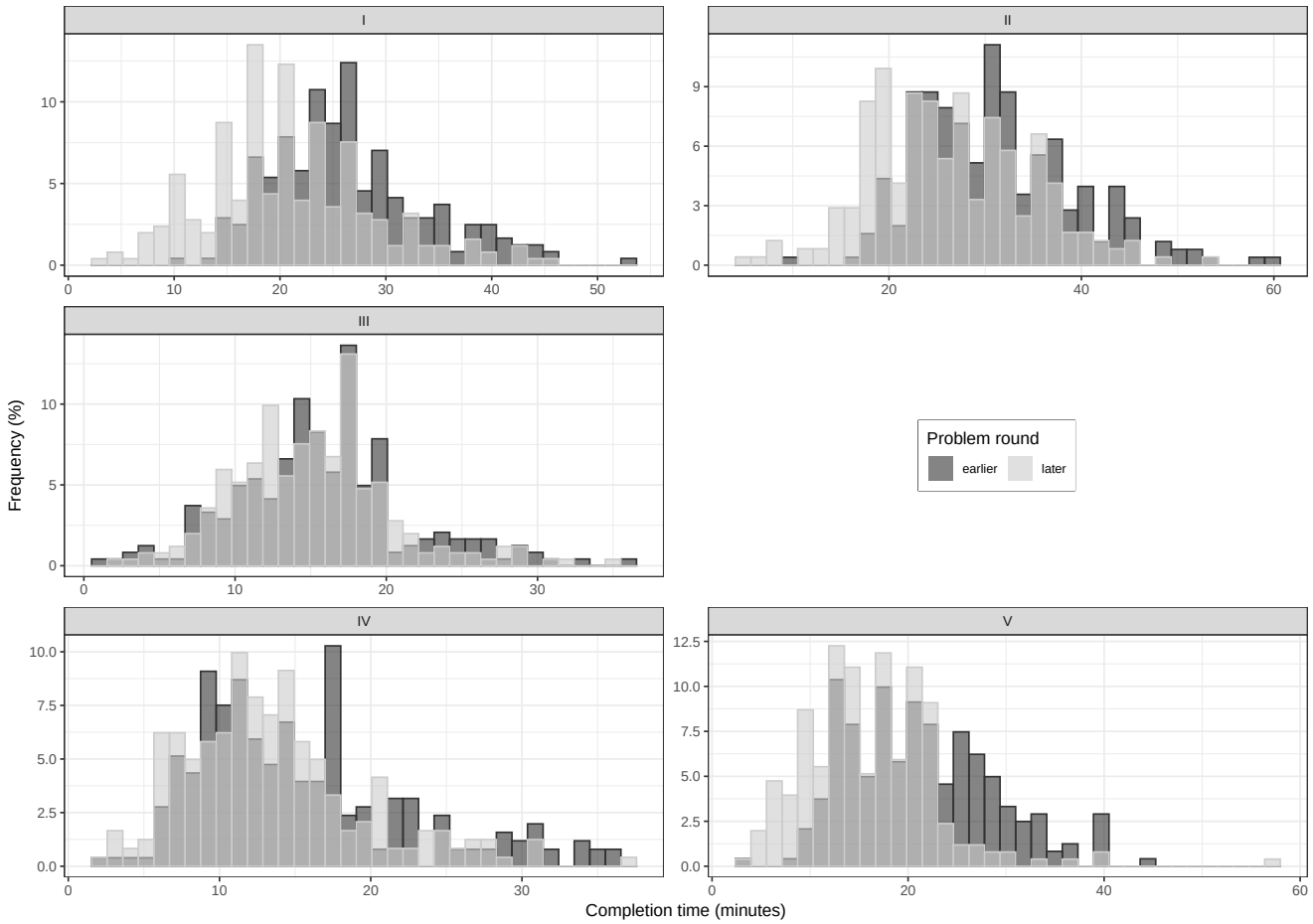

Figure 9: Frequencies of the completion time of each of the problems I, II, III, IV, and V. The "intermediate-gray" represents "earlier" when "later" has a higher frequency and it represents "later" when "earlier" has a higher frequency.

Figure 1 provided a visualization of individual data of all students. Figure 10 below provides scatter plots for the four groups of students separately (see Table 1). Each point represents one student's proportion of correct answers and completion time. As in Figure 1, we observe for all four groups that students that face a problem in a later round present a higher proportion of correct answers and shorter completion time. This is clearly reflected in the averages of the earlier and later group for problems I, II, IV, and V.

Group g3 presents better performance than the other groups in both dimensions. This is not very surprising given that g3 is the more demanding group that corresponds to the double bachelor's degree in Law – Business Management and Administration/Economics. The difference between the other more demanding group g4 (International Business Economics) and groups g1 and g2 seems less apparent.

Figure 10: Each row represents a different group of students. From left to right, panels describe correctness and completion time at the individual level for problems I, II, III, IV, and V separately. The triangles $\triangle$ and $\triangledown$ represent the averages of the earlier and later group, respectively.

Figure 11 below complements Table 5.

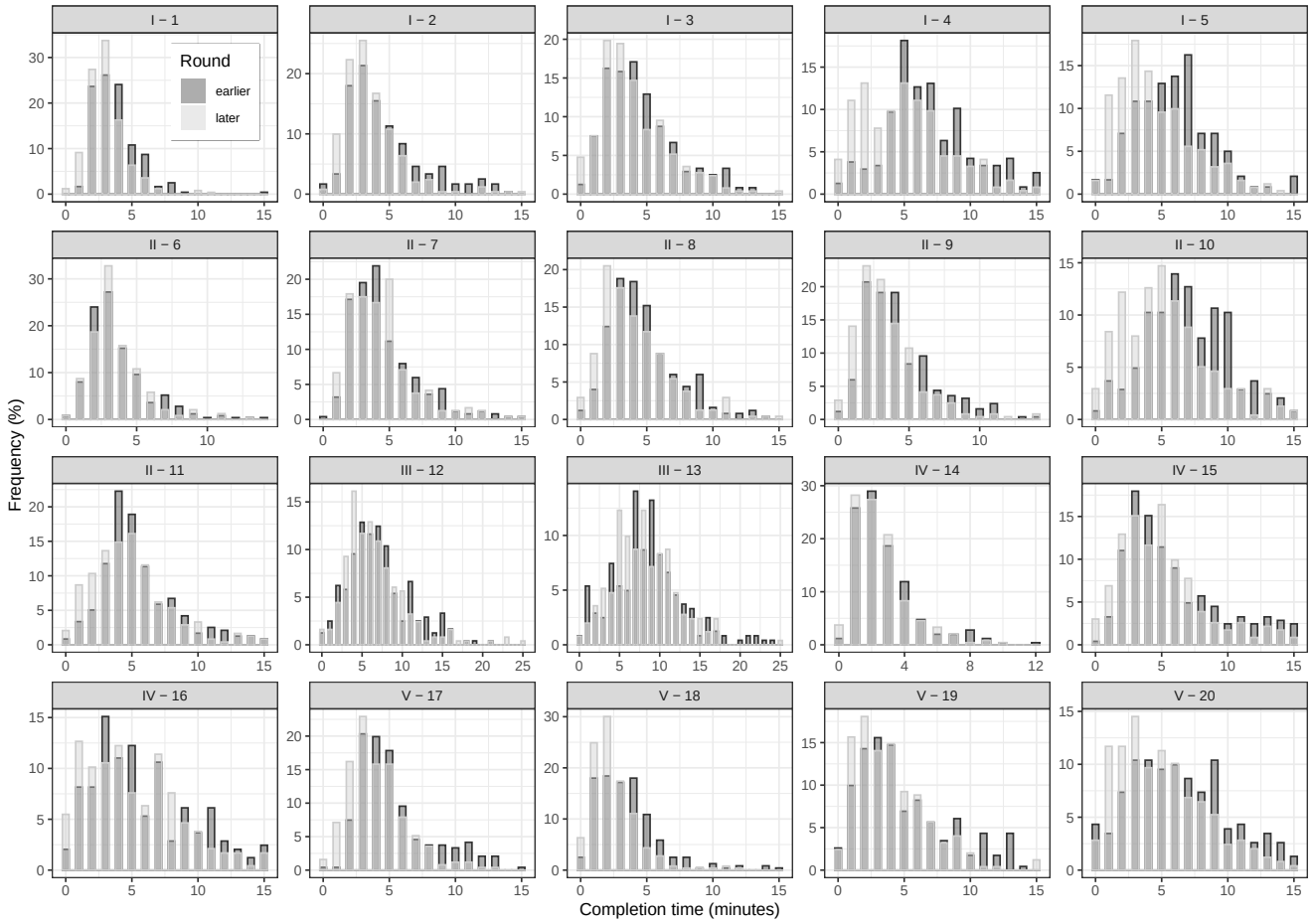

Figure 11: Frequencies of the completion time of each of the 20 questions separately. The "intermediate-gray" represents "earlier" when "later" has a higher frequency and it represents "later" when "earlier" has a higher frequency. For each of the questions in problems I, II, IV, and V we removed (at most) 3.4% of the (longer completion time) outliers. For each of the questions in problem III we removed (at most) 1% of the (longer completion time) outliers.

D Instant order effects

q	problem I - earlier					
	first		second		third	
3	0.835	(0.129)	0.930 (0.533)	(0.435)	0.880	
4	0.840	(0.507)	0.886 (0.219)	(0.690)	0.918	
5	0.707	(1.000)	0.699 (0.918)	(0.824)	0.727	
3	5.51	**(0.005)**	4.38 **(0.003)**	(0.913)	4.18	
4	8.56	**(0.012)**	7.17 **(0.000)**	**(0.002)**	5.32	
5	6.71	(0.081)	5.93 **(0.018)**	(0.400)	5.51	

q	problem I - later					
	first		second		third	
3	0.929	(0.958)	0.944 (1.000)	(0.925)	0.928	
4	0.893	(0.877)	0.874 (0.933)	(1.000)	0.877	
5	0.810	(0.461)	0.862 (0.471)	(1.000)	0.865	
3	4.15	(0.178)	3.64 (0.936)	(0.152)	4.15	
4	6.30	(0.130)	5.26 **(0.004)**	(0.162)	4.81	
5	5.29	**(0.033)**	4.32 **(0.004)**	(0.423)	3.98	

q	problem II - earlier		
	first		second
6	0.961	(0.936)	0.951
7	0.667	**(0.018)**	0.806
6	4.75	**(0.000)**	2.72
7	5.31	**(0.000)**	3.93

q	problem II - later		
	first		second
6	0.963	(0.662)	0.981
7	0.811	(0.939)	0.824
6	4.22	**(0.000)**	2.83
7	5.83	**(0.000)**	3.55

q	problem II - earlier					
	first		second		third	
8	0.433	(0.437)	0.507 (0.483)	(0.114)	0.371	
9	0.750	(0.276)	0.830 (0.653)	(0.663)	0.793	
10	0.791	(0.685)	0.753 (0.836)	(0.994)	0.765	
8	4.86	(0.491)	5.08 (0.155)	(0.0580)	4.59	
9	4.07	(0.491)	4.02 (0.937)	(0.4001)	3.95	
10	8.30	(0.085)	7.44 **(0.004)**	**(0.035)**	6.38	

q	problem II - later					
	first		second		third	
8	0.440	(0.842)	0.468 (0.069)	(0.151)	0.595	
9	0.866	(1.000)	0.864 (0.295)	(0.267)	0.931	
10	0.921	(0.568)	0.880 (0.859)	(0.854)	0.901	
8	4.90	(0.140)	4.02 (0.073)	(0.661)	3.89	
9	3.63	(0.694)	3.61 (0.162)	(0.076)	2.97	
10	5.34	(1.000)	5.65 (0.969)	(0.966)	5.46	

q	problem V - earlier		
	first		second
19	0.831	(0.494)	0.879
20	0.667	(0.084)	0.800
19	6.24	(0.086)	5.25
20	7.25	(0.509)	6.74

q	problem V - later		
	first		second
19	0.864	(0.260)	0.932
20	0.860	(0.495)	0.906
19	4.28	(0.197)	3.76
20	6.02	**(0.000)**	3.79

Table 8: Average correctness and completion time for each of the questions in the permutation groups [3,4,5], [6,7], [8,9,10], and [19,20]. The numbers with dark (light) gray background refer to correctness (completion time). The four tables on the left (right) hand side refer to the earlier (later) group. In brackets, the p-values which correspond to χ^2-tests of equal proportions (for correctness) and Mann-Whitney U tests (for completion time). In case of three possible positions of a question, the middle p-value corresponds to the test between the first and third positions. For instance, the p-value of 0.533 in the top-left corner corresponds to the χ^2-test that compares proportions 0.835 and 0.880.

Table 8 analyzes possible order effects for questions that were randomly permuted within a problem. For instance, it does not seem unlikely that in the group of students that are faced with problem I in round 1, students who face question 3 as their fifth question perform better (in terms of correctness and completion time) than the students who face question 3 as their third or fourth question. Table 8 summarizes our findings of this possible *instant order effect* for the four groups of questions [3, 4, 5], [6, 7], [8, 9, 10], and [19, 20].

Regarding correctness we make the following observations. First of all, looking at each pair of consecutive positions for each question, we observe that average correctness does *not* increase with the position of the question in 10 out of 32 cases. For instance, in problem I for the earlier group, the proportion of correct answers to question 3 is 0.930 when it appears second in the group [3,4,5] but reduces to 0.880 when it appears third. In fact, it can be observed that the only evident instance of an instant order effect for correctness is that of question 7 in problem II for the earlier group (proportions 0.667 and 0.806 with $p = 0.018$).

Our findings on completion time contrast with those on correctness. First of all, in 29 out of 32 cases of consecutive positions, completion time decreases with the position of the question.[22] Comparing the left hand side and the right hand side of Table 8, we observe that in the earlier group the instant order effect is stronger than in the later group. This is not surprising: all questions in the later round have already been faced by students in the earlier round and hence the particular position that a given question occupies in the later round can be expected to have less impact in terms of completion time (and correctness).

[22] In 10 of the 29 cases the decrease is statistically significant at the 5-percent level.

References

Alan, S., Ertac, S., and Gumren, M. (2020): "Cheating and incentives in a performance context: Evidence from a field experiment on children." *Journal of Economic Behavior and Organization*, 179: 681–701.

Bilen, E. and Matros, A. (2020): "Online cheating amid COVID-19." *Journal of Economic Behavior and Organization*, 182: 196–211.

Bretag, T. (2016): *Handbook of academic integrity.* Singapore: Springer.

Butler-Henderson, K. and Crawford, J. (2020): "A systematic review of online examinations: A pedagogical innovation for scalable authentication and integrity." *Computers & Education*, 159: 104024.

Cizek, G. J. and Wollack, J. A. (2016): *Handbook of quantitative methods for detecting cheating on tests.* New York: Routledge.

Crawford, V. P. and Sobel, J. (1982): "Strategic information transmission." *Econometrica*, 50(6): 1431–1451.

Dufwenberg, M. and Dufwenberg, M. A. (2018): "Lies in disguise – A theoretical analysis of cheating." *Journal of Economic Theory*, 175: 248–264.

Fischbacher, U. and Föllmi-Heusi, F. (2013): "Lies in disguise: An experimental study on cheating." *Journal of the European Economic Association*, 11(3): 525–547.

Frederick, S. (2005): "Cognitive reflection and decision making." *Journal of Economic Perspectives*, 19(4): 25–42.

Gneezy, U. (2005): "Deception: The role of consequences." *American Economic Review*, 95(1): 384–394.

Gurung, R. A., Wilhelm, T. M., and Filz, T. (2012): "Optimizing honor codes for online exam administration." *Ethics & Behavior*, 22(2): 158–162.

Holt, C. A. and Laury, S. K. (2002): "Risk aversion and incentive effects." *American Economic Review*, 92(5): 1644–1655.

Kartik, N. (2009): "Strategic communication with lying costs." *Review of Economic Studies*, 76(4): 1359–1395.

Martinelli, C., Parker, S. W., and Pérez-Gea, A. C. (2018): "Cheating and incentives: Learning from a policy experiment." *American Economic Journal: Economic Policy*, 10(1): 298–325.

McCabe, D. L., Butterfield, K. D., and Treviño, L. K. (2012): *Cheating in college: Why students do it and what educators can do about it.* Baltimore: The Johns Hopkins University Press.

McCabe, D. L., Treviño, L. K., and Butterfield, K. D. (2002): "Honor codes and other contextual influences on academic integrity: A replication and extension to modified honor code settings." *Research in Higher Education*, 43(3): 357–378.

Sánchez-Pagés, S. and Vorsatz, M. (2007): "An experimental study of truth-telling in a sender-receiver game." *Games and Economic Behavior*, 61(1): 86–112.

Sirota, M. and Juanchich, M. (2018): "Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the cognitive reflection test." *Behavior Research Methods*, 50: 2511–2522.

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.

Wollack, J. A. and Fremer, J. J. (2013): *Handbook of test security.* New York: Routledge.