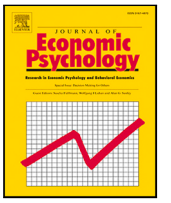
Journal of Economic Psychology 93 (2022) 102555

Contents lists available at ScienceDirect

Journal of Economic Psychology

journal homepage: www.elsevier.com/locate/joep



Brief report

Academic integrity in on-line exams: Evidence from a randomized field experiment☆

Flip Klijn [a,*,1], Mehdi Mdaghri Alaoui [b], Marc Vorsatz [c,2]

[a] Institute for Economic Analysis (CSIC) and Barcelona School of Economics, Campus UAB, 08193 Bellaterra (Barcelona), Spain
[b] Department of Economics and Business, Universitat Pompeu Fabra, C/ Ramon Trias Fargas 25, 08005 Barcelona, Spain
[c] Departamento de Análisis Económico, Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 11, 28040 Madrid, Spain

ARTICLE INFO

JEL classification:
A22
C93
D9
I21
I23

Keywords:
Field experiment
Academic integrity
Ethical code
On-line exam
Covid-19

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. First, 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 in the earlier round. Second, the average completion time of questions in the later round was 18.1% shorter than in the earlier round. Third, 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.

1. Introduction

Randomized field experiment

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. The main characteristics of the exam are as follows. First, students took the exam simultaneously. Second, the exam consisted of multiple-choice questions grouped into five problems. Each (but the third) 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, since all students faced exactly the same questions, the only difference between individual exams was the order of the questions. Fourth, backtracking was not possible, i.e., once a student moved to the next question, there was no possibility to go back to the previous question to change his/her answer to that question.

☆ We thank Rosemarie Nagel for useful comments and suggestions on an earlier version of the paper. We thank the editor and two anonymous reviewers for their valuable comments and suggestions which improved the paper. The raw data and R-scripts are publicly available at the Open Science Framework (OSF) address https://osf.io/yr6uc/.

* Corresponding author.
E-mail addresses: flip.klijn@iae.csic.es (F. Klijn), mehdi.mdaghri@upf.edu (M. Mdaghri Alaoui), mvorsatz@cee.uned.es (M. Vorsatz).

[1] He gratefully acknowledges financial support from AGAUR–Generalitat de Catalunya, Spain (2017-SGR-1359) and the Spanish Agencia Estatal de Investigación (AEI), Spain through grants ECO2017-88130-P and PID2020-114251GB-I00 (funded by MCIN/ AEI, Spain/10.13039/501100011033) and the Severo Ochoa Programme for Centres of Excellence in R&D, Spain (Barcelona School of Economics CEX2019-000915-S).

[2] He gratefully acknowledges financial support from the Spanish Ministry of Science and Innovation through grant PID2021-122919-NB-I00.

https://doi.org/10.1016/j.joep.2022.102555
Received 20 December 2021; Received in revised form 25 August 2022; Accepted 26 August 2022
Available online 5 September 2022

Main results

First, the students that received a given problem in the later round performed better in terms of higher correctness and shorter completion time. 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.

Related literature

Our study has similarities with four recent field studies. First, Martinelli et al. (2018) analyzed cheating in a governmental program on incentivized learning in mathematics in Mexico. Using statistical methods from the education measurement literature, the authors found that cheating is more prevalent under treatments that provide monetary incentives to students. Second, while Martinelli et al. (2018) concentrated on the correctness of the answers, Bilen and Matros (2021) considered 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 analyzed in detail the case of two students that present atypical time allocation to questions and extraordinary performance relative to midterm results. Third, Alan et al. (2020) presented a designed field experiment on cheating in a creative performance task for 720 elementary school children. They found that children with higher IQ and higher socioeconomic status have a higher likelihood of cheating. We do not find that the final exam grade is influenced by individual characteristics such as gender, risk aversion, and attention levels. Fourth, Vazquez et al. (2021) designed 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. Students whose exams were not proctored scored on average over 11% higher than those whose exams were proctored. 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.

Cheating behavior has also been studied extensively in the laboratory. Fischbacher and Föllmi-Heusi (2013) show in their seminal contribution on the dice-rolling experiment that a substantial proportion of subjects only partially misreport their private information to their advantage, maybe because of image concerns. Dufwenberg and Dufwenberg (2018) study the associated psychological game when players derive a disutility from being perceived to cheat. Recent contributions along this line, among others, are Pfattheicher et al. (2019), Siniver and Yaniv (2019), Huynh (2020), Fochmann et al. (2021), Clot et al. (2022), and Steinel et al. (2022). Pfattheicher et al. (2019) study the effects of personal traits and watching eyes on cheating in dice-rolling/coin-tossing experiments; Siniver and Yaniv (2019) show that a positive mood in the dice-rolling experiment is associated with more dishonest behavior; Huynh (2020) finds for Vietnamese college students that financial incentives do not affect cheating behavior in coin-tossing experiments unless subjects feel that they are facing a loss; Fochmann et al. (2021) analyze the effects of (correctly or incorrectly) prefilled forms in a tax compliance setting; Clot et al. (2022) employ experimental surveys to analyze how judges are affected by different reference points with respect to business wrongdoings; and Steinel et al. (2022) introduce uncertain outcomes into the dice-rolling paradigm.

Organization

In Section 2, we describe the field experiment and our hypotheses. In Section 3, we present our results. Finally, Section 4 applies the model of Martinelli et al. (2018) to quantify the cheating that is due to the different orders of the problems. The Online Appendix contains details about the course, screenshots of the final exam, subject pool information, additional analyses and figures, 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 2019–2020. The final exam was part of a continuous evaluation scheme which is discussed in more detail in Online Appendix A. The decision to run an on-line final exam was only taken by the end of the last lecture week (when a very strict lockdown due to Covid-19 started), approximately 10 days before the scheduled final exam. The final exam was programmed and executed in Moodle.

Design of final exam

All 494 students started the exam around the same time. The final exam consisted of *20 multiple-choice questions which were distributed over 5 problems* (see Online Appendix G). 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 whatsoever on their answers until two weeks after the exam.

The first screen (see Online Appendix A for screenshots of the final exam) 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, the number of problems, the number of points for a correct/incorrect/blank answer, and a reminder of the maximal duration (120 min). 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 a dice-rolling experiment with a monetary prize, for which we gave an additional 3 min.[3]

After clicking on the "continue" button at the bottom of the instructions, the first question 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 20 questions. However, problems and questions were permuted according to the scheme in Table 1.

Table 1
Distribution of the 20 questions (labeled $1, \ldots, 20$) over 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 after round 3.

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

For instance, students with version D 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). Students were not informed of the existence of different permutations of the exam. Online Appendix B provides more detailed information about our subject pool.

Hypotheses

Due to the very strict lockdown in Spain during the exam period it is unlikely that students worked together on the exam in the same physical space, but there were no impediments to on-line communication. We expect that (correct or incorrect) solutions/answers to any given question in the first round 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 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. In this definition, leaving a question unanswered is considered an incorrect answer. The total number of times a question was left unanswered is 343 or 3.47% of the total number of answers. The *average completion time* of a problem in a particular round is the average time taken for the problem by the students that were faced with the problem in that particular round.[4]

Hypothesis 1 (*Order Effect: Later Round Advantage*)**.** For each of 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. In particular, this suggests that potential "nervousness" that many students seem to experience at the beginning of any exam does not drive the later round advantage for the first two rounds (problems I and II). However, from an ex ante point of view, the result for problems IV and V 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.[5] Thus, the order of problems I and II should have no impact on the answers to problem III.

[3] The dice-rolling experiment is independent of the exam, in terms of both grade and time. It is analyzed in a separate study (currently work in progress). The instructions are in Online Appendix G (Question 21).

[4] A problem is considered completed by a student when he/she moves to the next problem. So, completion time of a problem includes the time used to read and think about any of its constituting questions and finally leaving it unanswered. The on-line platform measures time in minutes (where the minimum is 1 min).

[5] 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 started working on problem III around the same time.

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 other students (versions B and C). 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 present lower average correctness and longer average completion time afterwards than the other students.

3. Results

It is convenient to first investigate Hypothesis 3 because the no-result we obtain will allow us to pool the data afterwards. Reported *p*-values are two-sided throughout the whole study. Table 2 provides aggregate data of the last two rounds and focuses on the effect of the reminder. For instance, the first row of Table 2 compares the correctness and completion time of problem IV for students that face the problem in round 4: no reminder (version C) vs. reminder (version A).

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.

Problem	Round	Average correctness (proportion)				Average completion time (minutes)			
		no reminder	reminder	% *in*crease	*p*	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

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 (also see Figure 7 on problem IV and Figure 8 on problem V in Online Appendix F). There is no statistically significant comparison in Table 2. 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.

Next, 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 scatter plot in Fig. 1 visualizes the individual data. In Fig. 1 and *throughout the paper we 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). In each panel/problem, each point represents one student's proportion of correct answers and completion time.[6] Earlier (later) students are represented by circles ∘ (crosses +).

Fig. 1 supports Hypotheses 1 and 2. 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 ▽. Also, △ and ▽ are close to each other in problem III.

We statistically test for order effects between earlier and later students in Table 3. It is clear that for each of problems I and II, the earlier students achieve a higher average correctness and a shorter average completion time than the later students ($p < 0.001$ in the first two rows). Thus, we cannot reject Hypothesis 1. Figure 9 (on correctness) and 10 (on completion time) in Online Appendix F complement Table 3.

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 $26.4 + 26.3 = 52.7$ minutes to complete the two problems. Similarly, the students who had problem II in round 1 and problem I in round 2 (versions C and D) required a very similar time: on average $31.1 + 20.7 = 51.8$ minutes. In fact, we cannot reject the hypothesis that the two groups started problem III at the same time ($p = 0.435$, Mann–Whitney U test). This observation allows us to use the third row in Table 3 to see that Hypothesis 2 cannot be rejected: the difference in average correctness and average completion time between the earlier and later groups is not statistically significant.

[6] 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.

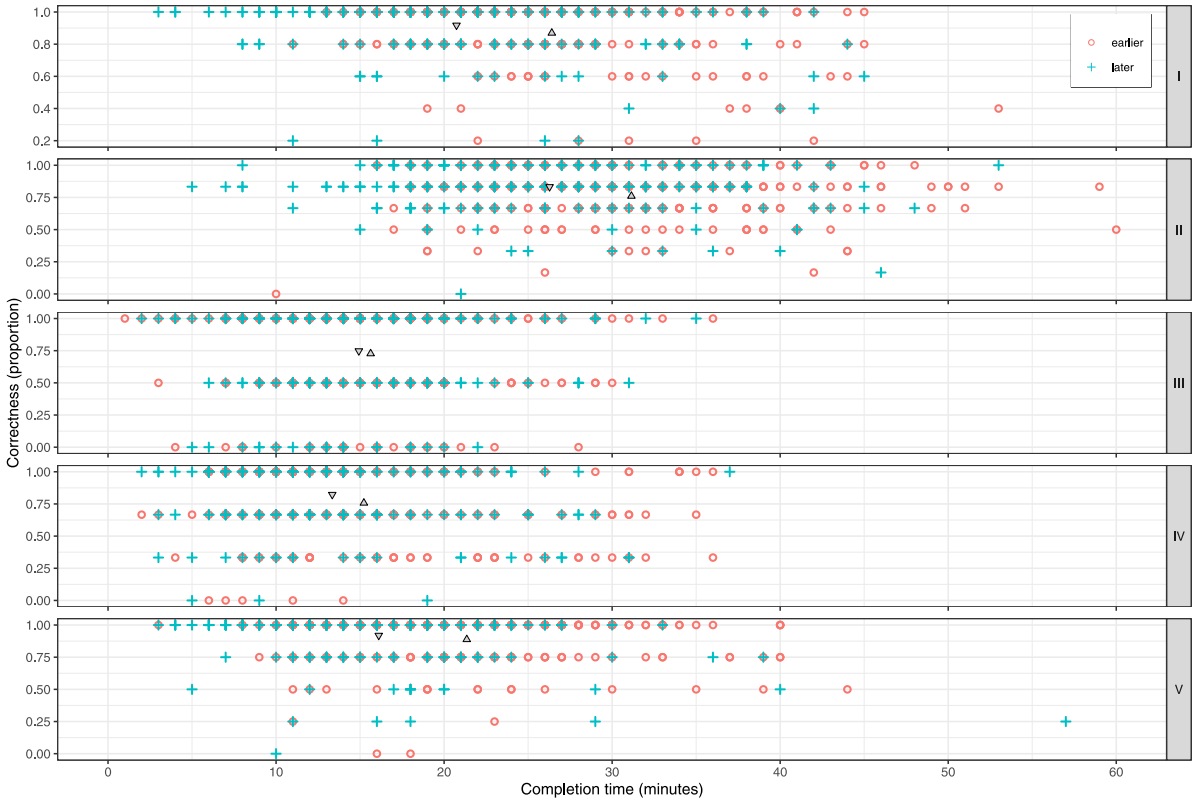


Fig. 1. From top to bottom, panels describe correctness and completion time at the individual level for problems I, II, III, IV, and V separately. The overlap of at least one circle ∘ and at least one cross + is visualized by a diamond-shaped form. The triangles △ and ▽ represent the averages of the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

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.

Problem	Average correctness (proportion)				Average completion time (minutes)			
	earlier	later	% *in*crease	*p*	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**

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 and completion time ($p < 0.01$ in the last two rows).

4. Concluding remarks

Quantification of "additional cheating"

It seems hard to provide a reliable estimate of the proportion of students that cheated at a particular question. However, we can apply the model of Martinelli et al. (2018) to gain some insights into the degree of *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.

Formally, for each $q \in \{1, 2, \ldots, 20\} \setminus \{12, 13\}$, let M_q be the set of possible answers to question q. Recall that each question has five fixed but randomly ordered answers. Let $m_{iq} \in M_q$ denote the answer of student i to question q. For each student i and each possible answer $m_q \in M_q$, we define $p_i(m_q) \equiv \texttt{Prob}(m_{iq} = m_q)$ as the probability that student i chooses answer m_q. Following the

Nominal Response Model of Bock (1972) it is assumed that

$$p_i(m_q) = \frac{exp(\zeta_{m_q} + \lambda_{m_q}\theta_i)}{\sum_{m'_q \in M_q} exp(\zeta_{m'_q} + \lambda_{m'_q}\theta_i)},$$

where $\lambda_{m_q}, \zeta_{m_q} \in \mathbb{R}$ are the distractors associated with $m_q \in M_q$ and $\theta_i \in \mathbb{R}$ is the (latent) ability of student i. We employ the R-package `mirt` to estimate the Nominal Response Model via maximum likelihood. The maximum likelihood estimator of $p_i(m_q)$ is denoted by $p_i^*(m_q)$.

The ω index of Wollack (1997) is then applied to assess whether there is an information flow from earlier round to later round students. Consider any problem $P \in \{\text{I, II, IV, V}\}$ and any ordered student pair (i, j), where i is the potential copier and j the potential information source. Let $h_{ij}(P)$ be the number of answers at problem P that coincide for students i and j. The ω index for the ordered student pair (i, j) with respect to P is defined as

$$\omega_{ij}(P) = \frac{h_{ij}(P) - \sum_{q \in P} p_i^*(m_{jq})}{\sqrt{\sum_{q \in P} p_i^*(m_{jq}) \cdot (1 - p_i^*(m_{jq}))}}.$$

Then, using an approximation of the ω index by means of the standard normal distribution function Φ, student i is classified as having copied from j on P if $1 - \Phi(\omega_{ij}(P)) \leq \alpha$. Here, α is the probability of making a false accusation. If there are n students who i could potentially copy from, then

$$r_i(\alpha, P) = \frac{\sum_j \mathbb{1}(1 - \Phi(\omega_{ij}(P)) \leq \alpha)}{n}$$

indicates how likely student i is to be classified for problem P as a copier when paired with a randomly selected student j.

We compare two scenarios to assess whether there is an information flow from earlier round to later round students. In the first scenario, both students in the student pair (i, j) are earlier round students. This benchmark controls for possible information flows within the group of earlier round students. In the second scenario, student i is a later round and student j is an earlier round student. So, if the average $r_i(\alpha, P)$, taken over all possible copiers i, is larger in the second scenario, later round students are more likely to copy from earlier round students than earlier round students among themselves. This would imply the existence of an information flow from earlier to later round students. The difference between the two scenarios can be taken as a measure of additional cheating.

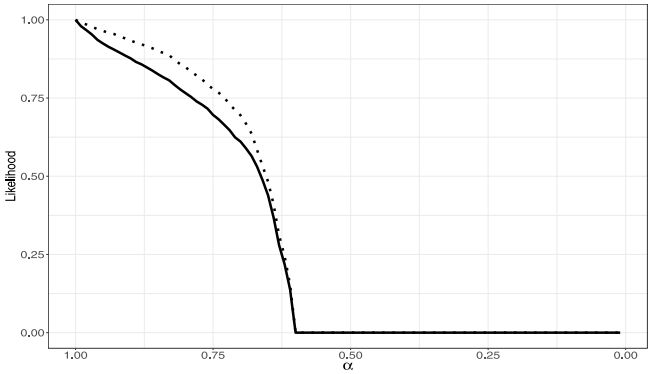


Fig. 2. Average $r_i(\alpha, P)$ aggregated over all problems $P \in \{I, II, IV, V\}$. Solid line: first scenario; dotted line: second scenario.

Fig. 2 depicts the average $r_i(\alpha, P)$ for $\alpha \in \{0.01, 0.02, \ldots, 1\}$. Note that we aggregate the data over all $P \in \{I, II, IV, V\}$. Most importantly, due to the small number of questions, students are never classified as copiers for $\alpha \leq 0.6$, which clearly limits the interpretability of the results. For $0.6 < \alpha < 1$, the curve of the second scenario always lies strictly above the one of the first scenario. Hence, consistent with our statistical analysis, there is a non-zero information flow from earlier to later round students. The maximum difference between the scenarios is 0.087, which corresponds to $\alpha = 0.71$. According to the model, additional cheating is thus bounded above by 0.087.

Randomization of questions and mitigation of cheating

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 seems a risky procedure for *on-line* exams, especially if there are no further measures to inhibit cheating. 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.

Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.joep.2022.102555.

References

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

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrica*, *37*(1), 29–51.

Clot, S., Grolleau, G., & Ibanez, L. (2022). A reference point bias in judging cheaters. *Journal of Economic Psychology*, *89*, Article 102485.

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

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

Fochmann, M., Müller, N., & Overesch, M. (2021). Less cheating? The effects of prefilled forms on compliance behavior. *Journal of Economic Psychology*, *83*, Article 102365.

Huynh, T. L. D. (2020). Replication: Cheating, loss aversion, and moral attitudes in Vietnam. *Journal of Economic Psychology*, *78*, Article 102277.

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

Pfattheicher, S., Schindler, S., & Nockur, L. (2019). On the impact of honesty-humility and a cue of being watched on cheating behavior. *Journal of Economic Psychology*, *71*, 159–174.

Siniver, E., & Yaniv, G. (2019). Optimism, pessimism, mood swings, and dishonest behavior. *Journal of Economic Psychology*, *72*, 54–63.

Steinel, W., Valtcheva, K., Gross, J., Celse, J., Max, S., & Shalvi, S. (2022). (Dis)honesty in the face of uncertain gains or losses. *Journal of Economic Psychology*, *90*, Article 102487.

Vazquez, J. J., Chiang, E. P., & 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*, Article 101653.

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, *21*, 307–320.