Detecting and Preventing Cheating in Exams: Evidence from a Field Experiment

Tobias Cagala, Ulrich Glogowsky, Johannes Rincke*

July 20, 2021

Abstract

This paper examines how to detect, document, and prevent plagiarism in exams. First, to identify and quantify plagiarism, we propose methods that compare similarities in multiple-choice answers between seat neighbors and non-neighbors. Second, we document cheating in undergraduate exams. Under baseline monitoring, at least 7.7% of the row-wise neighbor pairs plagiarized. Pairs composed of academically weaker students cheated more. Third, using a field experiment, we demonstrate that close monitoring eliminated cheating. By contrast, signing an honesty declaration doubled cheating relative to the control group. Complementary experiments suggest that the declaration backfired because it weakened the social norm of academic integrity.

Keywords: *academic cheating; close monitoring; honesty declaration; field experiment; randomization testing*

*Cagala: Deutsche Bundesbank (tobias.cagala@bundesbank.de); Glogowsky: Johannes Kepler University Linz (ulrich.glogowsky@jku.at); Rincke: University of Erlangen-Nuremberg (johannes.rincke@fau.de). We appreciate helpful comments and suggestions by Stefano DellaVigna, Emanuel Hansen, Johannes Hermle, Erzo Luttmer, Imran Rasul, Joel Slemrod, Christian Traxler, and seminar participants at various places. We are grateful for financial support from the Emerging Field Initiative of the University of Erlangen-Nuremberg. Ulrich Glogowsky also thanks Alan Auerbach and the German Research Foundation for supporting and funding his research stay at University of California (Berkely), during which this paper was partly written. The paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff. All errors are our own. This study is registered in the AEA RCT Registry (AEARCTR-0004612).

1 Introduction

Academic cheating is a wasteful illicit activity. It distorts the incentives for students to invest in their human capital, undermines the validity and usefulness of certificates as quality signals, and, hence, weakens the efficiency of the job-matching process (Spence, 1973). Due to these adverse effects, it is not only crucial to examine whether and how much students cheat but also vital to study how educators can promote academic integrity. However, because students try to conceal it, academic cheating is difficult to measure. Consequently, despite there being a long tradition of thinking about academic integrity (Parr, 1936; Drake, 1941; Bowers, 1964; Parnther, 2020), our knowledge of the incidence and nature of cheating is still incomplete.[1] Moreover, the measurement issue paired with missing exogenous variation has also prevented educators from thoroughly studying which countermeasures against academic cheating work and why.

In this paper, we examine how to detect, document, and prevent plagiarism in exams. We offer three contributions. First, we introduce strategies to identify and quantify plagiarism that compare similarities in multiple-choice answers of seat neighbors and non-neighbors. Second, exploiting these methods, we comprehensively document cheating in undergraduate exams at a German university. Particularly, we demonstrate that students copy answers from their seat neighbors, then we bound the amount of plagiarism, describe its spatial structure, and document the characteristics of the students who cheat. Third, we use a field experiment to evaluate whether two of the most popular countermeasures against academic cheating – close monitoring and the request to sign an honesty declaration – can promote academic integrity.[2] Because we implemented both countermeasures in a single setting, we can directly compare the interventions' effects.

Our evidence originates from written exams in introductory courses on economics and business administration. As discussed subsequently, these exams not only offer an environment that allows us to detect cheating but also one that enables us to conduct a clean and well-powered experiment. Moreover, in the baseline situation without experimental interventions, the students faced a setting with low levels of monitoring that is representative of many academic environments.

[1] Most of the evidence comes from survey data and suggests that academic cheating is widespread. For example, between 42% and 64% of participants stated that they had cheated in college at least once (Davis and Ludvigson, 1995). Furthermore, at Duke University, 21% of the surveyed undergraduates admitted to having cheated at least once a year (McCabe, 2005). Schab (1991) and Davis *et al.* (1992) provide further evidence. However, due to social-desirability biases or because subjects may fail to understand the principles of academic integrity (Power, 2009; Dee and Jacob, 2012), survey data can be problematic.

[2] Honesty declarations are widespread. For example, students frequently have to sign an honesty pledge when submitting assignments, term papers, or theses. Also, the honor-code system implies a well-known form of a pledge to academic integrity. According to the *U.S. News & World Report 2019*, the top 10 U.S. universities have an honor code or code of conduct that explicitly refers to academic integrity, and four out of the ten require undergraduate students to sign or pledge adherence to this code.

The ideal test for plagiarism would identify cheating by comparing the similarity in the seat neighbors' answers between a scenario in which students can cheat and a scenario in which it is impossible to cheat. A greater number of similarities identified in the scenario with cheating possibilities would indicate plagiarism. However, because the counterfactual scenario without cheating possibilities is unobservable, in practice, such a comparison is impossible. This complication sets the stage for our three-step approach to identify plagiarism. First, before the exam, we randomly assigned students to seats. Second, we approximate the similarities in the scenario without cheating possibilities by studying the answers of pairs of examinees who were not sitting next to each other (henceforth, counterfactual neighbors). Third, we compare the similarities in the answers of actual and counterfactual neighbor pairs to identify plagiarism. Due to seat randomization, the only reason why the similarity in the answers of actual and counterfactual pairs should differ is that the former could plagiarize from each other, while the latter could not. Hence, a comparison between actual and counterfactual neighbors enables us to identify plagiarism.

Exploiting this strategy, we comprehensively document plagiarism in undergraduate exams. Several insights emerge from the comparison of actual and counterfactual neighbors under weak baseline monitoring. First, the similarity in the answers of actual neighbor pairs is significantly higher than that in the answers of counterfactual neighbors. We conclude that the examinees plagiarized, even though they faced a no-cheating rule that the proctors announced before the exam. Second, we provide evidence on the spatial structure of cheating. The results suggest that the examinees copied answers from their row-wise neighbors. By contrast, we do not observe excess similarities for back-front neighbors or any other neighbor definition. Third, we demonstrate that most of the cheating happened in pairs in which at least one student had a below-than-median academic ability, measured by their high-school GPA. Fourth, regarding the amount of cheating, the lower-bound estimate for the share of cheating pairs is 7.7%. Additional evidence suggests that, on average, cheating pairs increased the number of shared answers by at most 45.6%. We conclude that, under baseline monitoring, cheating is widespread.

As a next step, our field experiment examines how to counteract the high levels of plagiarism. To do this, the experiment randomly assigned students to three groups. The control group implemented the university's standard exam conditions, consisting of weak (baseline) monitoring. We contrast this control group to two treatment conditions: a monitoring treatment and a signature treatment. The monitoring treatment created an environment of close monitoring by increasing the number of proctors per lecture hall. The signature treatment, meanwhile, implemented baseline monitoring but required students to sign an honesty declaration before the exam. Our main insights from the experiment are as follows. Under close monitoring, the answers of actual neighbors were not more (or less) similar than those of counterfactual neighbor pairs. Hence, in our exams, close monitoring eliminated all measurable traces of cheating. In sharp contrast to the findings on monitoring, our second countermeasure backfired and induced *more* cheating: in the signature treatment, the amount of cheating identified by our methods is twice as large as in the control group. Combining a second round of experiments with a post-exam survey, we also find that the examinees who had to sign the honesty declaration believed that cheating in exams was more common than students in the control group did. This result suggests that the signature treatment backfired because it weakened the perceived social norm of academic integrity.

Our paper contributes to several strands of literature. First, we extend the literature on the measurement of plagiarism that tests for cheating by examining whether the students' answers are unusually similar (see, e.g., Holland, 1996; Wollack, 1997, 2003, 2006; Wesolowsky, 2000; Sotaridona and Meijer, 2003; van der Linden and Sotaridona, 2006). Whereas the proposed approaches in this literature investigate if two suspected cheaters indeed plagiarized, we instead identify plagiarism in a large population of possible cheaters. To that end, we introduce the previously described identification approach of plagiarism. Closely related, Lin and Levitt (2020) also compare neighbors and non-neighbors to identify cheating. However, as they could only partially randomize students to seats, their methods rely on stronger assumptions.[3] In a different vein, we also extend the literature by providing techniques for bounding the amount of cheating, describing the characteristics of cheating pairs, and identifying the effects of countermeasures.

Second, we contribute to the emerging literature on the countermeasures against and the determinants of academic dishonesty. Lin and Levitt (2020) show that a bundled policy that jointly (a) prevents students from choosing a seat, (b) increases the number of proctors, and (c) shuffles multiple-choice answers can eliminate plagiarism. We, instead, identify the pure effect of close monitoring. Regarding honesty declarations, the literature offers little guidance on their usefulness to fight academic cheating, in spite of the fact that they are widespread. While some practitioners doubt that such declarations increase academic honesty (Cheung, 2012), descriptive work suggests that cheating tends to be lower at honor-code institutions that frequently utilize honesty declarations (Bowers, 1964; McCabe and Trevino, 1993; McCabe *et al.*, 2001). However, to our knowledge, the

[3] Lin and Levitt (2020) study a midterm and a final exam. In the midterm, examinees freely chose their seats. Hence, the analysis relies on the assumption that any excess similarity in the neighbors' answers reflects plagiarism and not the self-selection of similar students to neighboring seats. To test if this assumption holds, the final exam implements a clever design element: the authors record the seating chart under self-selection to seats but then randomly reseat students. This feature allows them to study if students who plan to sit next to each other give excessively similar answers, in line with the self-selection hypothesis. The evidence suggests that this is not the case, indicating that the excess similarities in the midterm reflect cheating. Our method is more direct: we randomize students to seats and, thereby, exclude self-selection biases by design. We also randomize students to treatments to study their effects.

causal effect of honesty declarations on academic cheating has yet to be analyzed. Other work demonstrates that classroom cheating responds to monetary incentives (Jacob and Levitt, 2003; Martinelli *et al.*, 2018), that anti-plagiarism tutorials reduce plagiarism in term papers (Dee and Jacob, 2012), and that social interactions (Lucifora and Tonello, 2015) and peer effects (Carrell *et al.*, 2008) amplify academic cheating. The evidence on social interactions and peer effects is in line with our finding that the adverse effect of an honesty declaration works through the perceived social norm of academic integrity.

Third, our paper is linked to a broader literature on fostering compliance with rules and norms in non-academic contexts. Following Becker (1968), a sizable empirical literature has studied how strategies involving monitoring and auditing affect compliance behaviors in the context of policing (Levitt, 1997; Di Tella and Schargrodsky, 2004), tax enforcement (Slemrod *et al.*, 2001; Kleven *et al.*, 2011), and fighting corruption (Olken, 2007; Ferraz and Finan, 2011).[4] Our finding that close monitoring in exams eliminates all measurable traces of cheating supports the consensus in this literature that monitoring and auditing are highly effective in promoting compliance. By contrast, the literature on how signed declarations affect compliance in non-academic settings is much smaller, and the evidence is mixed. To our knowledge, there are only two related field-experimental studies on the impact of honesty declarations from other contexts.[5] Shu *et al.* (2012) indicate that signing an honesty declaration placed at the beginning rather than at the end of an insurance self-report increases honesty. By contrast, the Behavioural Insights Team (2012) reports that moving an honesty declaration from the bottom to the top of a form used to apply for a tax discount likely increased fraud. This latter finding is well in line with emerging literature showing that well-intended interventions, such as trigger warnings (Jones *et al.*, 2020), can cause backfiring "boomerang effects" (see, e.g., the reviews of Miron and Brehm, 2006; Rains, 2013; Steindl *et al.*, 2015). In sum, it seems as if honesty declarations can produce diverging effects, depending on the context studied. Thus, the literature offers little guidance on whether educators should use honesty declarations to fight academic dishonesty.

The structure of the paper is as follows. Section 2 introduces our field experiment. Section 3 describes and applies our approaches to identify plagiarism and offers a detailed analysis of the nature of cheating. Section 4 examines how our treatments impact plagiarism and studies channels, and Section 5 concludes.

[4] Recent literature surveys include Chalfin and McCrary (2017) on policing, Slemrod (2019) on tax enforcement, and Olken and Pande (2012) on fighting corruption.

[5] There is also related evidence from the laboratory. Some papers suggest that interventions that confront individuals with morally charged information immediately before cheating decisions tend to increase honesty (Mazar *et al.*, 2008; Jacquemet *et al.*, 2019). However, one of the main findings supporting this view has recently come under attack: Verschuere *et al.* (2018) fail to replicate the finding of Mazar *et al.* (2008) that reminders of the Ten Commandments reduce misreporting (in 19 laboratories).

2 The Field Experiment

2.1 Background

Exams. We implemented the field experiment in two written, 60-minute undergraduate exams at the business school of a German university. The department's examination board and the responsible lecturers agreed to our interventions. The exams covered the courses "principles of economics" (first exam) and "principles of business administration" (second exam). The two exams took place on consecutive weeks. Moreover, each of the two exams included 30 multiple-choice problems. Each problem consisted of four statements, only one of which was correct. The examinees' task was to mark the correct statements on an answer sheet. All multiple-choice problems had the same weight, and incorrect answers did not impose a penalty.[6] Hence, a student's rational strategy was to mark a statement, even if she did not know the correct answer. Furthermore, the students answered all the problems in the same order.

The Setting's Benefits. Several aspects of the setting render it well suited to (a) identify cheating behavior and (b) implement a field experiment to test the effectiveness of close monitoring and honesty declarations. First, the exams included multiple-choice problems that, as discussed subsequently, allow us to detect cheating. Second, the exams were compulsory for students in their first semester and were part of the curriculum for a bachelor's degree. Because we focus on freshmen, students were unlikely to have noticed the changes in the examination conditions that we introduced with our treatments. Moreover, as many students had to take the exams, we are equipped with sufficient statistical power. Third, both exams took place across several lecture halls. This feature allows us to randomly allocate our treatments at the lecture-hall level, limiting spillovers between treatments. Fourth, the university did not have an honor code when we implemented the experiment, and monitoring in written exams was weak (see Subsection 2.2 for details). Furthermore, in the years before the experiment, the department did not request students to sign an honesty declaration before exams. These aspects provide us with a clean setting to study the impacts of our interventions. Fifth, given that we consider a low-enforcement environment, we focus on a setting that is representative of many academic contexts.

[6] We collected the exam data by scanning and electronically evaluating the multiple-choice answer sheets. This automated procedure ensures that the data are free from corrector bias and measurement error. We linked the exam data to administrative data on student characteristics.

2.2 Experimental Interventions

Randomization. The experiment randomly allocated students from two strata (gender and high-school GPA as a proxy for ability) to one of three experimental conditions: the control group, the signature treatment, and the monitoring treatment. To that end, we used a two-step randomization procedure. The first step randomized students from the strata to lecture halls; all the students in a hall received the same treatment.[7] The second step randomly assigned students to seats. As we will discuss later, our identification strategy of cheating critically relies on seat randomization.

Control Group. The control group implemented the department's standard examination conditions. Subsequently, we describe them, focusing on the setting's aspects that most likely affect the examinees' cheating decisions. The first likely driver of cheating behavior is the punishment in case of detection. According to the department's exam regulations, students who cheat (e.g., by copying answers from neighbors or using mobile phones) fail the exam. It is also part of the exam regulations that, before the exam, proctors announce standardized examination rules by reading them aloud. Figure B1 in Online Appendix B outlines these announcements. As part of the announcements, proctors highlight that cheating is prohibited and that detected cheaters fail the exam. The announcements also emphasize a list of actions that the administration classed as cheating attempts, including copying answers from neighbors, using unauthorized materials, and not switching off mobile phones. In the experiment, we made sure that proctors in all halls made the same announcements. As a result, it seems reasonable to assume that examinees in all halls similarly knew the consequences of cheating. Subsection 4.2 presents evidence in line with this notion.

A second essential element affecting cheating behavior is the monitoring level, as it influences the detection probability of cheating. Notably, the setting we study is one in which the baseline monitoring level is low. Up to 200 students take exams in lecture halls with up to 800 seats. However, only two to four university staff members (depending on the size of the hall) supervise the examinees. Moreover, the supervising staff have little incentive to monitor the examinees effectively. This is because proctors who intend to charge a student for cheating need to follow a complex protocol involving consultation with the department's examination board and considerable paperwork. Given these complications, it is no surprise that proctors rarely report cheating attempts. In fact, in the years before the experiment, no single student failed either of the two exams as a result of attempted cheating charges.

[7] We informed students before the exam in which lecture hall they would be seated. When arriving at the hall, they looked up their seat number on a list. Once all students took their seats, the proctors checked students' IDs and ensured they took their preassigned seats.

A third element that likely impacts plagiarism is the spatial distance between examinees, as it determines the students' ability to plagiarize answers. The seating arrangement in the experiment was as follows: students were sitting in every second row and every second column. Put differently, every examinee had to leave empty one seat to her left, one seat to her right, one seat in front of her, and one seat in her back. The row-wise distance between two students (1.2 meters, on average) was shorter than the column-wise distance (1.8 meters) or the diagonal distance (2.2 meters). This fact suggests that examinees could more easily copy answers from neighbors in the same row than from examinees sitting in the front or the back. Subsection 3.2 demonstrates that this is, in fact, the spatial pattern of cheating in our data.

Signature Treatment. The only difference between the control group and the signature treatment was that students in the signature treatment signed the following honesty declaration before the exam (see Figure B2 in Online Appendix B for details):

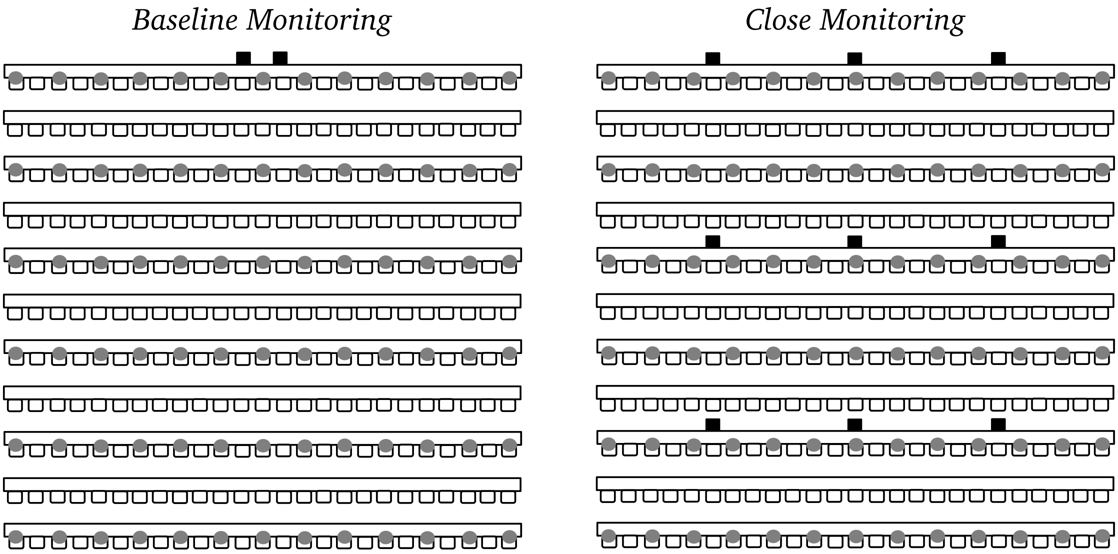
> *"I hereby declare that I will not use unauthorized materials during the exam. Furthermore, I declare neither to use unauthorized aid from other participants nor to give unauthorized aid to other participants."*

We printed this declaration on the cover sheet of the exam materials below a form that required examinees in all treatments to fill in their names and student IDs. This salient location was meant to direct the students' attention to the declaration immediately before the exam. Indeed, all the examinees in the signature treatment signed the declaration. Moreover, we gave students enough time to complete the form and sign the declaration before the exam. Examinees in all treatments, hence, had precisely 60 minutes to work on the exam itself.

Three further aspects of the signature treatment are essential. First, the declaration was neutral as opposed to morally loaded in that it did not refer to any ethical norm.[8] Second, the declaration did not introduce additional information regarding the no-cheating rule. Instead, the public announcements, which were identical across treatments, laid out the rule by stating that cheating was prohibited and by highlighting the consequences of cheating. Thus, the declaration neither varied the existence nor the content of the rule. By contrast, it aimed at changing the degree of commitment to an existing no-cheating rule relative to the control group. Third, the students in the control group did not have to sign the form on the cover sheet. We chose this design element as we are interested in the effects of a *signed* honesty declaration.

[8] We used a neutral declaration because practitioners frequently use this type, both in education-related settings and beyond. We do not claim that morally loaded declarations would have had the same effects.

Figure 1: Monitoring Conditions in the Field Experiment



Notes: This figure is a stylized illustration of baseline monitoring (control group and signature treatment) and close monitoring (monitoring treatment). Gray dots represent examinees; black squares represent proctors. The average monitoring intensities were 44.2 examinees per proctor under baseline monitoring and 8.4 examinees per proctor under close monitoring.

Monitoring Treatment. Our monitoring treatment heavily increased the monitoring intensity compared to the control group and the signature treatment. Specifically, the monitoring treatment implemented close monitoring by allocating additional proctors to the lecture halls such that, on average, one proctor monitored only 8.4 examinees. By contrast, the baseline monitoring level in the control group and the signature treatment was much lower: in these groups, there were, on average, 44.2 examinees per proctor. To control the monitoring level, we also ensured that proctors in all halls remained at specific predefined spots throughout the exam. In the monitoring treatment, these spots were evenly distributed all over the halls. Instead, in the control group and the signature treatment, proctors took positions in the hall's front. Figure 1 sketches the hall setups under baseline and close monitoring. Importantly, the other aspects of the monitoring treatment were identical to the control group. Thus, the examinees in the monitoring treatment did not sign an honesty declaration.

2.3 Further Details and Sampling

Further Details. We took several further steps to guarantee that all examination conditions other than the treatment variations were constant across the lecture halls. First, we harmonized proctor behavior across halls. To that end, the supervising staff had to follow a scripted schedule. The script included the exact wording of all the announcements to be made before and after the exam. Second, we equalized the monitoring conditions

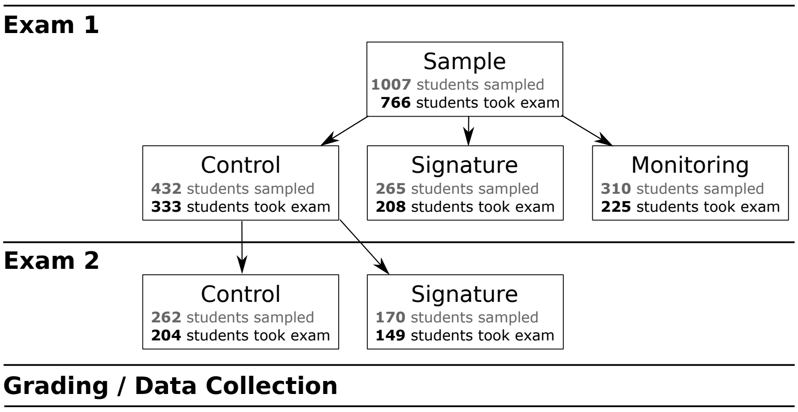
Table 1: Balancing Checks

	Exam 1					Exam 2		
	Control	Signature	Monitoring	Difference Signature–Control	Difference Monitoring–Control	Control	Signature	Difference Signature–Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Gender (Female = 1)	0.54	0.56	0.50	0.02	-0.04	0.53	0.53	0.00
				(0.04)	(0.04)			(0.05)
High-School GPA	2.47	2.48	2.50	0.02	0.03	2.48	2.50	0.01
				(0.05)	(0.05)			(0.06)
Math Proficiency	0.75	0.73	0.73	-0.02	-0.02	0.75	0.74	-0.01
				(-0.01)	(-0.01)			(0.02)
Field of Study (Econ. & Sociology = 1)	0.07	0.06	0.09	-0.01	0.02	0.08	0.08	0.00
				(0.02)	(0.02)			(0.03)
Age	19.6	19.6	19.6	-0.04	-0.03	19.7	19.5	-0.15
				(0.10)	(0.10)			(0.11)
Bavaria	0.81	0.83	0.84	0.02	0.02	0.80	0.83	0.03
				(0.03)	(0.03)			(0.04)
Number of Observations	333	208	225			204	149	

Notes: This table shows balancing checks for both exams covered in the field experiment. Columns (1) to (3) report treatment-specific means for Exam 1. Column (4) shows the difference in means between signature and control with standard errors in parentheses. Column (5) reports the difference in means between monitoring and control. Columns (6) to (8) report means and the difference in means for Exam 2. High-School GPA is the grade point average from high school (criterion for university admission), ranging from 1.0 (outstanding) to 4.0 (pass). Math Proficiency is obtained from a university math exam taken prior to the exams studied in the experiment. The proficiency score gives the percentage of total points the student obtained in the math test. Field of Study is a dummy for students with a major in Economics & Sociology, the reference group being students enrolled in Economics and Business Administration. Bavaria is a dummy for students who finished high school in Bavaria. Gender and High-School GPA were used for stratification.

within all close-monitoring halls and also those within all baseline-monitoring halls. For example, we ensured that the actual examinee-per-proctor ratios in the halls were identical to the planned ones.[9] There were also no asymmetries in the number of empty seats between the treatments that would have altered the cheating opportunities of the participating students in an ex-ante, unknown way. Third, we ensured that all the conditions related to the treatment interventions were unobservable to examinees before the beginning of the exam. In particular, the proctors distributed the exam materials and entered the halls only after all the students took their preassigned seats. As a result, on-the-spot decisions regarding whether or not to take part in the exam should be uncorrelated with the treatment assignment. Indeed, we do not find systematic differences in the students' observable characteristics between the control and treatment groups (see Table 1).

Figure 2: Overview of Field-Experimental Design



Notes: This figure visualizes the experimental design. We implemented the field experiment in two written exams. Exam 1 comprised a control group, the signature treatment, and the monitoring treatment. Students assigned to the control group in Exam 1 were also sampled for the intervention in Exam 2, comprising a control group and a signature treatment group. The figure indicates, for each treatment, the number of students assigned to the respective treatment group and the number of students who took the exam. Differences between the two numbers are due to the fact that students could postpone participation to later semesters.

Sampling. Figure 2 presents an overview of the sampling scheme. Our overall sample consisted of 1007 students eligible to take the exams. In the first exam, we randomly assigned 432 students to the control group, 265 to the signature treatment, and 310 to the monitoring treatment. The show-up rates did not vary significantly between the treatment groups and ranged between 73% and 78%.[10] Ultimately, 766 examinees took the

[9] Due to local examination conditions, students could withdraw from the exam up until the exam day. To prevent a no-show effect on the examinee-per-proctor ratios, we overbooked lecture halls when randomly allocating students to treatments. Due to the overbooking procedure, some students could not be seated in their preassigned hall. We reseated those students to additional halls that were not part of the experiment.

[10] Two rules that are typical for German universities explain the low show-up rates. First, students can decide when to sit exams. The only requirement is that, after three terms, they must have passed 10 out of

first exam: 333 in the control group, 208 in the signature treatment, and 225 in the monitoring treatment. The sampling frame for the second exam used only the 432 students from the first exam's control group. Hereby, we ensured that all the considered students shared a similar treatment history. We did not implement the monitoring treatment in the second exam. Consequently, we assigned the 432 students to the control group or the signature treatment of the second exam. Of the sampled students, 353 took the second exam (control: 204; signature: 149). Notably, if not stated otherwise, the following evidence relies on a sample that pools both exams (i.e., we analyze both exams jointly). We do not find any evidence that the results for the first and second exams are statistically different from each other.

3 Detecting Cheating in Exams

3.1 Basic Idea of Tests for Cheating

Basic Idea. Our identification approach starts from the idea that plagiarism leaves detectable traces in the data. If examinees plagiarize, the similarities in seat neighbors' answers are higher than in a counterfactual scenario without cheating.[11] In practice, however, the counterfactual scenario is not observable, and we must find ways to approximate how the similarities would look in the absence of cheating. For that purpose, we propose two tests: a *non-parametric randomization test* and a *regression-based test*. Both tests build on a similar core: they approximate the counterfactual scenario without cheating by creating many counterfactual neighbor pairs consisting of students who were not sitting side by side. Given the spatial distance, counterfactual neighbors could not plagiarize from each other, providing us with an approximation of the counterfactual scenario. The tests then explore if the similarity in the answers of actual neighbors is statistically higher than the similarity in the responses of counterfactual neighbors.

Identifying Assumption. The identifying assumption of our tests is that plagiarism is the only systematic reason why the similarity in the answers of actual neighbors is different from that of counterfactual neighbors. Following Manski's (1993) framework for identifying social effects, the assumption holds under two conditions (see, e.g., Manski, 2000; Blume *et al.*, 2011; Herbst and Mas, 2015, for discussions). First, the composition of both types of pairs needs to be identical. For example, neighbors and non-neighbors must, on average, have similar individual characteristics. Second, both types of pairs must

the 12 courses that, according to the curriculum, optimally should be taken in the first two terms. Second, the university does not punish "no shows."

[11] Figure B3 in Appendix B exemplifies the spatial patterns in the answers by showing examinees' answers to one multiple-choice problem in one particular control-group hall.

face an identical institutional environment during the exam. The examination conditions, for example, need to be identical for both types of pairs.

Ensuring that the Assumption Holds. We took two steps to ensure that both conditions were met. First, to guarantee that there were no systematic differences in the composition of pairs, we randomly assigned individuals to seats and formed counterfactual pairs randomly. Hence, we followed the standard approach in the social-effects literature and exploited randomization schemes to allocate individuals to groups within which interactions may occur (see, e.g., Sacerdote, 2001; Falk and Ichino, 2006; Kremer and Levy, 2008; Guryan *et al.*, 2009). Second, to ensure that actual and counterfactual neighbors faced the same institutional environment, we only use non-neighbors who sat in the same lecture hall to construct counterfactual neighbor pairs. This decision nets out lecture-hall effects. Note that our approach identifies cheating in the form of plagiarism only. Other forms of cheating (like, for instance, using crib sheets) stay undetected by our methods. We, therefore, likely understate the actual incidence of cheating. However, our conclusions will hold as long as the treatment effects are uncorrelated with the cheating technology.

3.2 Prevalence of Cheating in Exams

In the following, we identify cheating behavior under baseline monitoring using a spatial randomization test and a complementary regression-based test.

Randomization Test: Method. Randomization testing goes back to Fisher (1922) and is a standard inference tool in the analysis of experiments. The key characteristic of randomization tests is that, instead of relying on a theoretical distribution, they compare a test statistic to a null distribution obtained from the data by resampling.[12] Applied to our context, we consider a test statistic that measures the similarity of neighbors' answers. We then test if this measure is unusually high compared to its null distribution (i.e., the distribution in the absence of cheating), which we obtain by a resampling procedure that constructs counterfactual neighbors. Along these lines, our procedure allows us to test against the null hypothesis that the similarities in actual neighbors' answers are not different from those in counterfactual neighbors' answers.

More specifically, our baseline testing algorithm consists of four steps:

[12] Randomization testing is widespread. Researchers rely on it to calculate inference for treatment effects (Rosenbaum, 2002; Duflo *et al.*, 2008). Other papers use randomization schemes to test how outcomes of individuals are connected. For example, Falk and Ichino (2006) use an approach similar to ours to identify peer effects in co-workers' productivity.

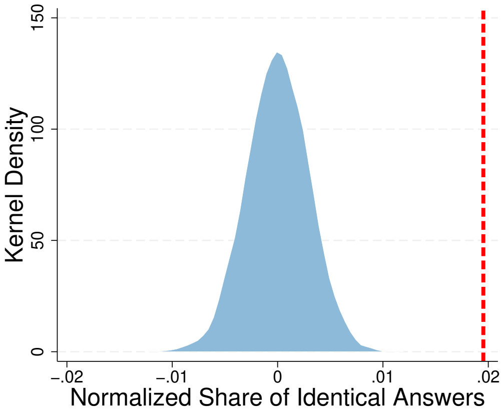
1. We calculate the share of all multiple-choice problems $\widehat{s}_{i,i-1}$ that an examinee i and her left neighbor $i-1$ in the same row r answered identically (correct or incorrect). We do the same for i and her right neighbor $i+1$ to derive $\widehat{s}_{i,i+1}$ and compute the test statistic as:
$$\widehat{\Delta} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}(\widehat{s}_{i,i-1}+\widehat{s}_{i,i+1}),$$
where N is the number of examinees under baseline monitoring.
2. We create counterfactual neighbor pairs by randomly reassigning examinees within halls to seats. Our reassignment procedure ensures that the counterfactual pairs do not consist of two examinees who were actually sitting in the same row. This feature avoids that the null distribution picks up similarities induced by the row-wise copying of answers. We then compute the similarity measure for the counterfactual neighbors, $\widehat{\Delta}_{C,m=1}$.
3. We repeat the second step M times. Hereby, we generate a distribution of $\widehat{\Delta}_{C,m}$ with the values $m = 1, ..., M$, mean $\widehat{\mu}_{\widehat{\Delta}_C}$, and standard deviation $\widehat{\sigma}_{\widehat{\Delta}_C}$. Under our identifying assumption, this distribution corresponds to the distribution of the test statistic under the null hypothesis of no cheating.
4. We calculate the p-value of a two-tailed test as twice the probability that a draw from this distribution exceeds $\widehat{\Delta}$.

Our baseline test aims to identify plagiarism from direct neighbors sitting in the same row. If examinees copied answers from other individuals farther away, the test would underreject its null hypothesis. We, however, find no indications of other forms of cheating (see paragraph "robustness checks").

Randomization Test: Baseline Results. Figure 3 reports the results of our baseline randomization test ($M = 5000$). It relies on data from both exams and both baseline-monitoring treatments. For ease of exposition, the figure reports mean-centered values (i.e., it shows $\widehat{\Delta}-\widehat{\mu}_{\widehat{\Delta}_C}$ and $\widehat{\Delta}_{C,m}-\widehat{\mu}_{\widehat{\Delta}_C}$). This type of normalization allows us to interpret the test statistic intuitively as the extent to which the share of identical answers among actual neighbors differs from the expected share for counterfactual neighbors (in percentage points). The vertical line depicts the mean-centered test statistic. The bell-shaped curves represent the mean-centered counterfactual distributions under the null hypothesis of no cheating.

The main finding of Figure 3 is that, under baseline monitoring, the similarity in the answers of actual neighbors is excessively high compared to the counterfactual distribution. The test statistic indicates that the share of actual neighbors' identical answers is almost two percentage points higher than the expected share for counterfactual neighbors. Moreover, the test statistic lies in the far right tail of the counterfactual distribution,

Figure 3: Cheating Under Baseline Monitoring



Notes: This figure shows the results for our randomization tests, considering the baseline-monitoring sample. The vertical line represents the test statistic derived from the actual seating arrangement. The bell-shaped curve plots the mean-centered null distribution based on Epanechnikov kernels. We obtain $p < 0.001$ (two-tailed test with Bonferroni correction).

and we can, consequently, clearly reject the null hypothesis of no above-normal similarity in the answers of actual neighbors (p-value < 0.001).[13] This finding is the first piece of direct evidence that under baseline monitoring, examinees copied answers from their direct row-wise neighbors.

Randomization Test: Further Results and Robustness Checks. Online Appendix B presents several additional analyses and robustness checks. First, as mentioned, we investigate if examinees plagiarized from peers sitting farther away. Figures B6 to B9 present evidence for a variety of alternative specifications (see Panels A). The figures suggest that, in our context, only direct row-wise neighbors have plagiarized from each other. Put differently, we do not observe excess similarities for back-front neighbors or any other neighbor definition. Second, we probe the robustness of our results to the resampling scheme. For example, we resample individuals within treatments (i.e., also across halls) instead of within halls (see Figures B6 to B9).[14] Following Cagala *et al.* (2019), we also present tests that do not exclude counterfactual neighbors sitting in the same row (see Figures B8 and B9). Our findings are robust. Third, Panel C in Figures B4 and B5 tests for cheating under close monitoring. In line with the hypothesis that close monitoring nullifies or at least sharply reduces cheating, we cannot reject the null hypothesis that the similarities

[13] Including the specifications that are part of our robustness checks, we use six different neighbor definitions to test for cheating. To guard against spurious findings from multiple testing, we employ a conservative Bonferroni adjustment to correct the reported p-values.

[14] We prefer to resample individuals within halls because this resampling scheme controls for potential hall effects and, hence, decreases the probability of false positives. The flipside is that this scheme potentially increases the likelihood of false negatives: if present, the counterfactual distribution would pick up plagiarism across rows. A randomization scheme that also resamples individuals across halls alleviates this potential problem (by construction, examinees in different halls could not plagiarize from each other).

in actual neighbors' answers are identical to those in counterfactual neighbors' answers ($p > 0.999$).[15] Section 4 explores the role of close monitoring in more detail.

Regression-Based Test: Method. While the randomization tests allow us to test for the presence of cheating without relying on parametric assumptions, they offer relatively little flexibility. For example, they do not provide a simple way to study effect heterogeneity and, hence, the characteristics of cheating pairs. In the next step, we introduce a regression-based test that offers this flexibility. Before turning to the models that allow us to study the nature of cheating, we introduce a baseline model that, in a similar vein to the randomization tests, solely tests for the presence of cheating.

The model, again, rests on the idea of using counterfactual neighbors to identify cheating between actual row-wise neighbors. Specifically, we estimate the following linear probability model with OLS:

$$Y_{mp} = \beta_0 + \beta_1 N_p + u_{mp}, \tag{1}$$

where Y_{mp} takes a value of one if both students of a pair p gave the same (correct or incorrect) answer to a particular multiple-choice problem m. Note that p can represent actual and counterfactual pairs. Furthermore, N_p indicates whether a pair of students consisted of actual neighbors sitting next to each other in the same row ($N_p = 1$) or not ($N_p = 0$). The estimated coefficient $\widehat{\beta}_0$ measures the probability that counterfactual neighbor pairs give an identical answer. Instead, under random assignment to seats, $\widehat{\beta}_1$ is a consistent reduced-form estimate of the average effect of being an actual neighbor pair (instead of a counterfactual pair) on the probability of identical answers. We call this estimate the average neighbor effect (*ANE*). An *ANE* significantly greater than zero indicates cheating.

Two further details of our regression-based approach are worth noting. First, it identifies the *ANE* using the same counterfactual neighbors as the randomization tests. To do this, we define counterfactual neighbors as pairs of students in the same hall who, however, sat in different rows. Second, we base statistical inference on a hall-level wild-cluster-bootstrap procedure (Cameron *et al.*, 2008). Notably, this method of inference likely underrejects the null hypothesis if (as in our case) only a few clusters are treated or untreated (see, e.g., MacKinnon and Webb, 2017). Hence, our approach is conservative.

Regression-Based Test: Results. Table 2 reports estimates of the ANE, focusing on all lecture halls with baseline monitoring and both exams. Column (1) shows uncon-

[15] This finding also suggests that our tests, indeed, identify plagiarism in exams. Otherwise, we would not necessarily expect that the monitoring treatment reduces the similarities in actual neighbors' answers.

ditional estimates. The estimated ANE is positive and significantly different from zero ($p < 0.001$). The regression, thus, replicates the finding of the randomization test that students cheated under baseline monitoring. The effect is also sizable: compared to the 57.5% probability that counterfactual neighbors shared an identical answer, the probability among actual neighbors is 2.02 percentage points, or 3.5%, higher. Column (2) adds our complete set of control variables to model (1). Specifically, it controls for multiple-choice fixed effects, hall fixed effects, and two types of pair-specific variables: control variables for gender combinations (a female-female dummy and a male-male dummy) and controls for high-school grade combinations (grade indicators for the better and worse student as well as interactions). Because we randomly assigned examinees to seats, there is no reason to expect the controls to affect the average neighbor effect. Indeed, the estimate in Column (2) is only slightly different from that in Column (1).

Table 2: Average Neighbor Effect Under Baseline Monitoring

	Dependent Variable: Indicator for Identical Answer	
	(1) Unconditional Estimate	(2) All Controls
Actual Neighbors	0.0202 [0.0002]	0.0188 [0.0004]
Multiple Choice FE	No	Yes
Hall FE	No	Yes
Pair Controls	No	Yes
Mean for Counterfactual Neighbors	0.577	
Number of Clusters	8	
Number of Observations	1,121,034	

Notes: This table reports estimates of the average neighbor effect on the probability that two paired students provide identical (correct or incorrect) answers under baseline monitoring. The estimates rely on linear probability models. The specifications define counterfactual neighbors as pairs of students in the same hall who did not sit next to each other. Column (1) presents the unconditional estimates. Column (2) adds controls (multiple-choice fixed effects, hall fixed effects, indicators for gender combinations, and indicators for high-school grade combinations). All specifications also include an exam dummy. Wild-cluster-bootstrap p-values in [brackets].

Regression-Based Test: Further Results and Robustness Checks. Again, we present additional results and robustness checks in Online Appendix A. Table A1 considers identical correct and incorrect answers separately. The examinees seem to have plagiarized correct and incorrect answers, although the estimates for incorrect answers are more precise. Furthermore, Table A2 demonstrates that, when using non-neighbors sitting in the same row as counterfactual neighbors, the results remain essentially unchanged.[16]

[16] In a previous version, we used this counterfactual definition to derive our results (Cagala *et al.*, 2019). The benefit of this alternative approach is that it indirectly controls for row effects as it compares counter-

3.3 Nature of Cheating

The purpose of this subsection is to explore the nature of cheating. The results demonstrate (a) how cheating correlates with the students' ability and (b) how it shifts the distribution of identical answers. The distributional analysis also allows us to provide a lower bound for the share of cheaters.

Grade Heterogeneity: Method. The previous subsection established that actual neighbors shared a suspiciously high number of similar answers under baseline monitoring. We expect to detect especially strong traces of plagiarism if at least one of the two students of a pair is an academically weaker student. Intuitively, weaker students should be less able to succeed in the exam, increasing their need to cheat.

To test this hypothesis, we estimate an extended version of the model (1) that allows the neighbor effect to vary in the pairs' ability composition. Specifically, we approximate an examinee's academic ability by her final high-school GPA and estimate the model:

$$Y_{mp} = \beta_0 + \beta_1 N_p + \beta_2 H_p + \beta_3 M_p + \beta_4 H_p \times N_p + \beta_5 M_p \times N_p + u_{mp}. \qquad (2)$$

As apparent, the model interacts the binary indicator for actual neighbors N_p with two dummy variables, measuring the ability composition of pair p. The first dummy, H_p, indicates if both students of pair p performed better in high school than the median student. The second one, M_p, is a dummy variable for pairs in which one student performed better and the other one worse than the median student. Crucially, the interacted structure of the model allows the neighbor effect to vary across three types of pairs: "high-ability pairs" consisting of two above-median-ability students (i.e., $H_p = 1$), "low-ability pairs" being composed of two below-median-ability students (i.e., $M_p = 0$ and $H_p = 0$), and "mixed pairs" (i.e., $M_p = 1$). The OLS estimates of β_1 measure the *ANE* for low-ability pairs. By contrast, the estimates of β_4 and β_5 capture deviations from this baseline neighbor effect for high-ability pairs and mixed pairs.[17]

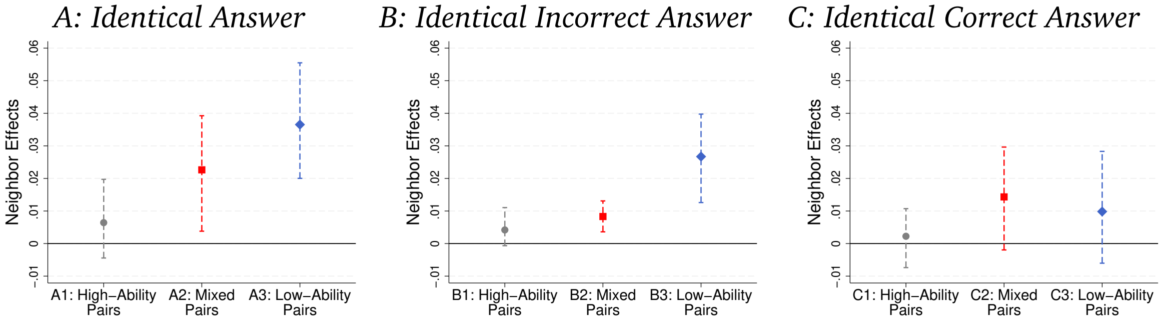
Grade Heterogeneity: Results. Figure 4 graphically decomposes the average neighbor effects, considering the pooled baseline-monitoring sample. To construct the figure, we estimate three versions of the model (2) that use different outcome variables: a dummy indicating if both students of pair p answered question m (a) identical (Panel A), (b) identical and incorrect (Panel B), or (c) identical and correct (Panel C). Moreover, Figure

factual neighbors and actual neighbors within rows. The drawback is that the estimated neighbor effects might be too small due to cascades of cheating within rows which confound the counterfactual. All the results of our paper are widely unchanged when applying the previous estimation strategy.

[17] If N_p is randomized, the estimates of β_1, β_4, and β_5 are unbiased and consistent. By contrast, the estimates of β_2 and β_3 pick up potential correlations between the grade variables and u_{mp}.

4 separately depicts the average neighbor effects for high-ability pairs (gray circles),[18] mixed pairs (red squares),[19] and low-ability pairs (blue diamonds).[20] It also shows wild-cluster-bootstrap 95% confidence intervals.

Figure 4: Grade Heterogeneity in the Average Neighbor Effect



Notes: This figure examines how the students' ability (proxied by high-school GPA) relates to their cheating behavior under baseline monitoring. To construct this figure, we estimate the effects of being a pair of actual neighbors on the probability that two students give identical answers (Panel *A*), identical incorrect answers (Panel *B*), or identical correct answers (Panel *C*). Crucially, we allow the effects to vary in whether both students (gray circles), one student (red squares), or none of the students of pair p (blue diamond) performed better in high school than the median student. All specifications include an exam dummy and derive the 95% confidence bands by a wild-cluster-bootstrap procedure.

Panels *A* to *C*, indeed, show substantial heterogeneity in the average neighbor effect. First, considering all identical answers as an outcome, the average neighbor effect for high-ability pairs is relatively small and not significantly different from zero (estimate *A1*). The same result applies if we separately consider identical incorrect (*B1*) or identical correct answers (*C1*). These results suggest the absence of significant cheating among examinees of above-median ability. Second, in line with the hypothesis that below-median-ability students cheated, we find significant neighbor effects if one of the two examinees performed worse in high school than the median student (*A2*). The point estimates for identical incorrect (*B2*) and identical correct answers (*C2*) are relatively similar, suggesting that cheaters cannot distinguish between true and false answers. However, our estimate for identical correct answers is less precise and only significant at the 10% level. Third, we report stronger above-normal spatial correlations in identical answers for low-ability pairs (*A3*). The fact that the point estimate for low-ability pairs is higher than for mixed pairs is in line with the interpretations that (a) mainly low-ability students cheated and (b) both students of low-ability pairs copied from each other (bidirectional cheating).[21] The reason is that we identify cheating at the pair level. Hence, we expect

[18] The gray circles depict $\widehat{E}[Y_{mp}|N_p = 1, H_p = 1, M_p = 0] - \widehat{E}[Y_{mp}|N_p = 0, H_p = 1, M_p = 0] = \widehat{\beta}_1 + \widehat{\beta}_4$, where $\widehat{E}[\cdot]$ denotes a conditional average computed on the sample.

[19] The red squares show $\widehat{E}[Y_{mp}|N_p = 1, H_p = 0, M_p = 1] - \widehat{E}[Y_{mp}|N_p = 0, H_p = 0, M_p = 1] = \widehat{\beta}_1 + \widehat{\beta}_5$.

[20] The blue diamonds plot $\widehat{E}[Y_{mp}|N_p = 1, H_p = 0, M_p = 0] - \widehat{E}[Y_{mp}|N_p = 0, H_p = 0, M_p = 0] = \widehat{\beta}_1$.

[21] We cannot reject the hypothesis that the ANE for low-ability pairs is twice the size of mixed pairs.

larger neighbor effects for pairs in which both examinees plagiarized from each other than for pairs with only one cheater. Fourth, low-ability pairs rather tended to copy wrong answers from each other (compare *B3* and *C3*). This result is intuitive: students should only plagiarize questions that they cannot answer themselves. Crucially, two low-ability neighbors likely fail to solve the same (difficult) questions, resulting in more plagiarism of incorrect than correct answers.[22]

Grade Heterogeneity: Discussion and Conclusions. To sum up, the first conclusion of Figure 4 is that mainly the below-median-ability examinees seemed to have cheated. This finding is in line with studies based on self-reported data (Genereux and McLeod, 1995; McCabe and Trevino, 1997). One straightforward explanation is that more able students have less need to cheat because they can succeed in the exam without help. Moreover, the high-ability individuals might even take a competitive stance and shield their answers to prevent other examinees from copying. Low-ability students, instead, might be more willing to allow copying (or they even trade their answers). In a different vein, high-ability students might also cheat less because they hold different views on academic integrity. The second conclusion of Figure 4 is that, in line with Lin and Levitt (2020), cheating leaves more easily identifiable traces in jointly incorrect than jointly correct answers. First, the precision of the estimated neighbor effects for jointly incorrect answers is much higher. Second, the effects for identical incorrect answers are also larger for low-ability pairs, which also simplifies identification for this subgroup. We conclude that identical incorrect answers introduce less noise and reflect a larger part of cheating for low-ability pairs. Hence, they are the more powerful indicator of plagiarism in our context. In the following, we consequently use identical incorrect answers as our primary outcome variable.

Grade Heterogeneity: Robustness Checks. Online Appendix B presents several robustness checks. Figure B10 includes multiple-choice fixed effects, lecture-hall fixed effects, and pair controls. Figure B11 defines the grade indicators based on the performance of the mean student instead of the median student. Both checks leave the results essentially unchanged.

Distributional Analysis: Method. By definition, pairs composed of cheaters share a higher number of identical answers than non-cheating pairs. Hence, starting from a scenario without cheating, we expect that plagiarism shifts mass in the distribution of

[22] We do not expect the same effect for mixed pairs. High-ability students more likely solve questions that are unsolvable for low-ability students, rendering it more likely that low-ability students copy correct answers.

identical answers from lower to higher numbers. We test this hypothesis under baseline monitoring.

To study the distributional impacts of cheating, we proceed in three steps. The first step plots the distributions of identical incorrect answers for counterfactual and actual neighbor pairs and compares them. In line with cheating, the distribution for actual neighbors should feature more mass in the right tail than the distribution for counterfactual neighbors. In a second step, we test if the distributions for actual and counterfactual neighbor pairs are, indeed, statistically different from each other. We use a non-parametric Wilcoxon signed-rank test that accounts for potential correlations at the hall level.[23] The third step directly compares the distributions visually by calculating and plotting the difference between the actual and counterfactual distribution for each value of X, where X counts the number of identical incorrect answers. We label the resulting X-specific differences "neighbor effects on the distributions" (NED^X).

Distributional Analysis: Results. Figure 5 presents the results from the distributional analysis for the pooled baseline-monitoring sample. The solid red line in Panel A depicts the fraction of actual neighbor pairs that share X identical answers, f^X. The dashed blue line shows the respective fraction for counterfactual neighbor pairs, $\widetilde{f}^X$. Panel B depicts the corresponding neighbor effects, $NED^X = f^X - \widetilde{f}^X$.[24] For visibility, the figure focuses on X-values between 0 and 8. Figure B13 in the Appendix shows the distributions up to $X = 12$, which is the maximum of identical incorrect answers given by a pair.

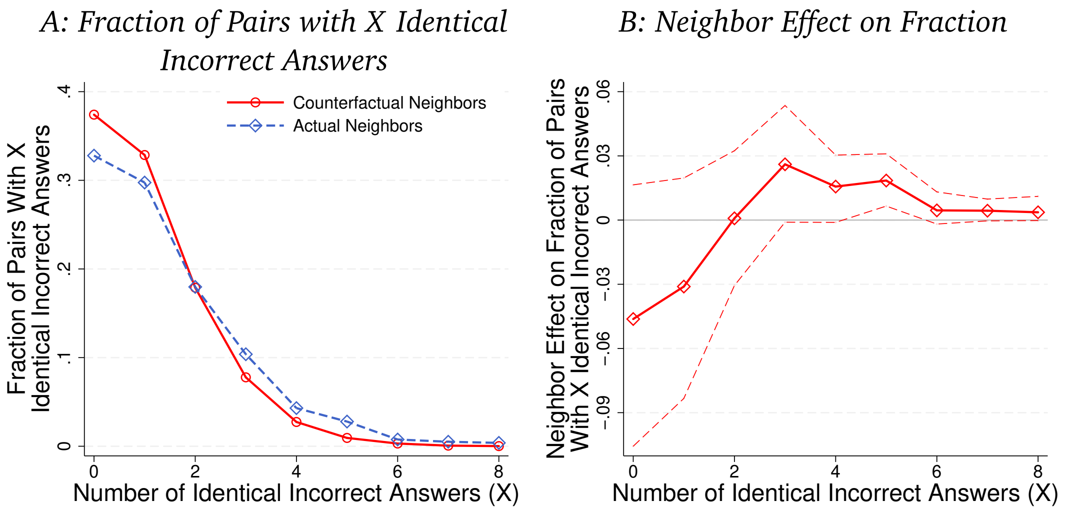
The key insight from the analysis is that, relative to the counterfactual distribution, some mass in the distribution for actual neighbor pairs seems to be shifted to the right. Indeed, the Wilcoxon signed-rank test rejects its null hypothesis that both distributions are the same ($p = 0.028$). Panel A of Figure 5 demonstrates the underlying shift to the right visually. Specifically, the fraction of actual neighbors who share less than two identical incorrect answers is lower than that for counterfactual neighbors. This missing mass appears to be shifted to higher X-values: the mass for actual neighbors above $X = 1$ exceeds that for counterfactual neighbors (Panel A), resulting in positive neighbor effects (Panel B). The NED^X peaks at $X = 3$ and converges towards zero for $X > 5$.[25] Given that we do not observe much more excess mass in the right tail of the distribution for

[23] The test collapses the data to the hall level. It then compares the two related samples, "counterfactual neighbors" and "actual neighbors," to assess whether their population mean ranks differ.

[24] Note that we can also estimate the neighbor effect with regressions. Specifically, for each X, we may estimate: $Y_p^X = \beta_0^X + \beta_1^X N_p + u_p \ \forall \ X = 0, ..., 30$, where Y_p^X is a binary outcome variable, indicating if the two students of pair p gave precisely X identical incorrect answers ($Y_p^X = 1$) or not ($Y_p^X = 0$). In these regressions, β_0^X reflects the fraction of counterfactual neighbor pairs that share X identical answers, $\widetilde{f}^X$. By contrast, β_1^X identifies the average neighbor effect on Y_p^X, $f^X - \widetilde{f}^X$.

[25] The statistical power increases mechanically in X, explaining the wider CIs for lower X values.

Figure 5: Shift in the Distribution of Identical Incorrect Answers



Notes: This figure shows how cheating shifts the distribution of identical incorrect answers. Panel *A* depicts the distribution of identical incorrect answers for counterfactual neighbor pairs (solid red line) and actual neighbor pairs (dashed blue line). Panel *B* shows the corresponding neighbor effects, which are the X-specific differences between the actual and counterfactual distribution shown in Panel *A*. To construct the 95% confidence bands in Panel *B*, we estimate the model described in Footnote 24 and employ our standard wild-cluster-bootstrap procedure.

actual neighbor pairs, we conclude that, on average, examinees copied a limited number of answers from their neighbors rather than copying entire exams.

3.4 Amount of Cheating

Next, we quantify the amount of cheating. In a first step, we derive a lower bound for the share of neighbor pairs that plagiarized. Building on this estimate, in a second step, we then provide back-of-the-envelope calculations for the average number of answers copied by cheating pairs.

Share of Cheaters: Method. Our lower-bound estimate measures the share of all neighbor pairs in which one or both students cheated. It follows from comparing the distributions of identical incorrect answers between actual and counterfactual pairs (see Figure 5). The main complication of such an aggregate comparison is that it does not reveal the cheating behavior of every single pair of actual neighbors. Hence, it does not allow us to pin down the fraction of cheaters precisely. However, as Appendix C discusses in more detail, the comparison allows us to bound the share of cheaters. To that end, we employ a distribution-based two-step procedure: first, we obtain the set of total aggregated neighbor effects $TNE = \{TNE^1, TNE^2, ..., TNE^{29}\}$ with $TNE^{X^*} = \sum_{X=X^*}^{30} NED^X$.[26] Each

[26] Hence, TNE^{X^*} is the total excess mass of the actual over the counterfactual distribution above X^*.

of the total aggregated neighbor effects TNE^1 to TNE^{29} reflects a subset of cheaters. In particular, TNE^{X^*} measures the share of neighbor pairs that, by copying from each other, increased their number of identical incorrect answers from less than X^* to X^* or more.[27] Second, we analyze which of the 29 subsets of cheaters, given by the aggregated neighbor effects, is the largest and define the lower bound as the largest of those subsets.[28] Thus, our strategy maximizes the share of cheating pairs identifiable with distributional analyses.

Share of Cheaters: Results. Figure B14 in Online Appendix B presents the estimates for the total aggregated neighbor effects TNE^1 to TNE^{12}. The estimated total aggregated neighbor effect is the largest for $X^* = 2$, with a value of $TNE^2 = 7.7\%$. This value defines our lower-bound estimate for the share of neighbor pairs that comprise at least one cheating student. Intuitively, it implies that the subset of cheaters that increased their number of identical incorrect answers from less than two to two or more answers amounts to 7.7% of all pairs. Notably, our lower bound for the share of cheaters is well in line with Lin and Levitt (2020), who identify cheating behavior in at least 10% of their examinees. We conclude that cheating occurred in a sizable fraction of neighbor pairs.

Share of Cheaters: Robustness Checks. There are two reasons why the presented estimate for the share of cheaters, TNE^2, identifies a lower bound: first, the analysis exploits only incorrect answers, our less noisy outcome. However, if some pairs systematically copied only correct answers, TNE^2 underestimates the share of cheating pairs. Second, as the aggregated analysis cannot identify if a particular pair cheated or not, it potentially misses some forms of cheating. For example, TNE^2 neglects those cheating pairs that, without cheating, would already share more than two identical incorrect answers.[29] To test the robustness of our results to these limitations, Online Appendix D extends our randomization tests. The extended test identifies cheating at the pair level and relies on all identical (correct and incorrect) answers. The results suggest that the distribution-based lower bound provides a reasonable estimate of the percentage of cheating pairs.

[27] To see this, recall that our identifying assumption states that plagiarism is the only systematic reason why the similarity in answers differs between actual and counterfactual neighbors. Moreover, by definition, cheating increases the similarities. Hence, any excess mass in the distribution for actual neighbors at value X reflects the share of pairs that, by cheating, increased the number of identical incorrect answers from less than X to X. The TNE^{X^*} aggregates the excess mass for $X \geq X^*$ and, therefore, measures the share of pairs that increased the number of identical incorrect answers from less than X^* to X^* or more.

[28] Intuitively, we, thus, analyze which fraction is the largest: the fraction of pairs that, by cheating, increase the number of identical answers from less than one to one or more (measured by TNE^1), the fraction of pairs that increase it from less than two to two or more (measured by TNE^2), and so on.

[29] Consider, for example, a pair of cheaters that, due to cheating, shares four instead of three identical answers. This type of behavior would leave our lower bound $TNE^2 = \sum_{X=2}^{30} (f^X - \widetilde{f}^X)$ unaffected. The reason is that the increase in f^4 fully compensates for the decrease in f^3 so that TNE^2 remains constant.

According to the extended estimates, between eight and ten percent of the pairs engaged in cheating.

Amount of Cheating Among Cheaters. Finally, we discuss our back-of-the-envelope calculations for the average number of answers copied by cheating pairs. To that end, recall that cheating increased the probability that actual neighbor pairs shared an identical (correct or incorrect) answer by 2.02 percentage points (see Column (1) in Table 2). Transformed to the exam level consisting of 30 multiple-choice problems, actual neighbor pairs, hence, shared $0.0202 \cdot 30 = 0.61$ additional identical answers compared to counterfactual neighbor pairs. However, not all of the actual neighbor pairs plagiarized. Using the lower bound estimate for the share of cheating neighbor pairs of 7.7%, we can approximate an upper bound for the average number of copied (correct and incorrect) answers among cheating pairs as $0.61/0.077 = 7.9$.[30] Thus, on average, cheating pairs seem to have plagiarized at most 7.9 identical answers, which corresponds to 26.2% of the 30 problems. As the average number of identical (correct and incorrect) answers among counterfactual neighbor pairs was 17.3, we can also state that, on average, cheating pairs increased the number of shared answers by at most 45.6%.

4 Preventing Cheating in Exams

4.1 Effectiveness of Close Monitoring and Honesty Declarations

Method. To examine if the signature and monitoring treatments can prevent cheating, we extend the model (1) such that the neighbor effect can vary by treatment. Specifically, we estimate the following regression with OLS:

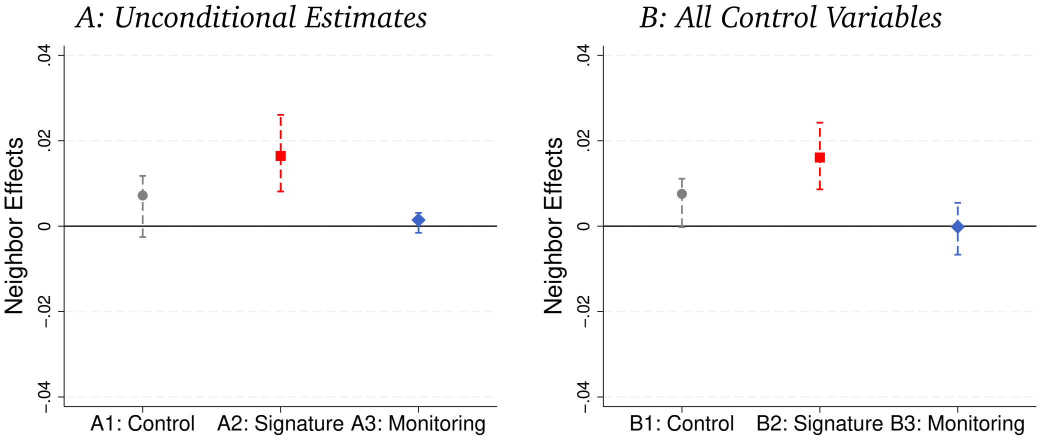
$$Y_{mp} = \beta_0 + \beta_1 N_p + \beta_2 ST_p + \beta_3 MT_p + \beta_4 ST_p \times N_p + \beta_5 MT_p \times N_p + u_{mp}, \qquad (3)$$

where Y_{mp} is a binary indicator for identical incorrect answers, N_p indicates if pair p consisted of actual neighbors, ST_p indicates whether the students of pair p received the signature treatment, and MT_p is a binary indicator for the monitoring treatment. In this equation, the estimate $\widehat{\beta}_0$ measures the probability that counterfactual neighbors in the control group give an identical incorrect answer. By contrast, $\widehat{\beta}_2$ and $\widehat{\beta}_3$ show if and to what extent the signature and monitoring treatments change this probability for counterfactual neighbors. Moreover, under random treatment assignment, the estimate $\widehat{\beta}_1$

[30] For simplicity, the calculation uses the estimate for the share of cheaters derived from the distributions of identical incorrect answers. If we, instead, exploit the randomization estimate presented in Appendix D that relies on identical correct and incorrect answers, the result is almost identical (copied answers: 7.6). The reason is that both estimation strategies for the share of cheaters deliver similar results.

identifies the neighbor effect in the control group. Consequently, $\widehat{\beta}_4$ and $\widehat{\beta}_5$ reflect deviations from this baseline neighbor effect in the signature and monitoring treatment. Again, we employ the conservative wild-cluster-bootstrap procedure for statistical inference.

Figure 6: Treatment Heterogeneity in the Average Neighbor Effect



Notes: This figure shows the average neighbor effect on identical incorrect answers for the control group (gray circles), the signature treatment (red squares), and the monitoring treatment (blue diamonds). To construct this figure, we estimate model (3) and, based on this model, predict the treatment-specific neighbor effects. Panel *A* presents the unconditional estimates. Panel *B* adds the complete set of control variables to our model (hall fixed effects, multiple-choice fixed effects, indicators for gender combinations, and indicators for high-school grade combinations). All specifications also include an exam dummy and derive the 95% confidence bands by a wild-cluster-bootstrap procedure.

Results. Table 3 demonstrates the impacts of the signature and monitoring treatments on cheating, pooling the data over both exams. Again, Column (1) presents unconditional estimates, and Column (2) includes the complete set of control variables. The lecture-hall effects in Column (2) fully absorb the baseline effects for the signature and monitoring treatments. In addition to the table, Figure 6 shows the corresponding treatment-specific neighbor effects for the models without (Panel *A*) and with control variables (Panel *B*).[31]

We present four main results. First, in line with the notion that students in the control group cheated, the point estimates of the average neighbor effects are positive and amount to 0.0072 and 0.0076. Although employing a conservative inference method, we can reject the null hypothesis of no effect at the 10% level (without controls) and 6% level (with controls). The estimates imply that the probability that actual neighbors shared an identical incorrect answer was between 0.72 percentage points (≈ 19.5%) and 0.76 percentage points (≈ 20.5%) higher than the 3.7% probability that counterfactual neighbors shared such an answer.

[31] The gray circles depict $\widehat{\beta}_1$, the red squares show $\widehat{\beta}_1 + \widehat{\beta}_4$, and the blue diamonds plot $\widehat{\beta}_1 + \widehat{\beta}_5$.

Table 3: Responses to the Signature and Monitoring Treatments

	Dependent Variable: Indicator for Identical Incorrect Answer	
	(1) Unconditional Estimates	(2) All Controls
Signature	0.0004 [0.8928]	
Monitoring	−0.0007 [0.8217]	
Actual Neighbors	0.0072 [0.0917]	0.0076 [0.0545]
Signature × Actual Neighbors	0.0092 [0.0469]	0.0085 [0.0524]
Monitoring × Actual Neighbors	−0.0057 [0.0861]	−0.0077 [0.0215]
Multiple Choice FE	No	Yes
Hall FE	No	Yes
Pair Controls	No	Yes
Mean for Counterfactual Neighbors	0.037	
Number of Clusters	11	
Number of Observations	1,412,787	

Notes: This table demonstrates how the treatments change the likelihood that two paired students provide identical incorrect answers. The estimates rely on linear probability models. Column (1) presents the unconditional estimates. Column (2) adds controls (multiple-choice fixed effects, hall fixed effects, indicators for gender combinations, and indicators for high-school grade combinations). All specifications also include an exam dummy. Wild-cluster-bootstrap p-values in [brackets].

Second, compared to the control group, the signature treatment significantly and substantially increased the probability that actual neighbors shared an identical answer. More specifically, the neighbor effects in the signature treatment are more than twice the size of the effects in the control group (see Figure 6). The corresponding interaction effects *Signature* × *Actual Neighbors* lie between 0.0085 and 0.0092 with $p \leq 0.0524$. These estimates imply that actual neighbors in the signature treatment increased the probability of sharing an identical incorrect answer (relative to counterfactual neighbors) by 0.85 and 0.92 percentage points more than actual neighbors in the control group. Moreover, we cannot reject the hypothesis that the interaction effects are equal to the average neighbor effect in the control group ($0.7762 \leq p \leq 0.8817$). Taken together, the findings suggest that the signature treatment has at least doubled the amount of cheating compared to the control condition.

Third, the evidence suggests that close monitoring eliminated cheating. Figure 6 demonstrates that the neighbor effects in the monitoring treatment are not only precisely estimated but also very close to zero (see the estimates $A3$ and $B3$ that amount to 0.0015 and 0.0005). Table 3 shows that this is because the coefficients of the interaction term

Monitoring × *Actual Neighbors* are negative and (in absolute values) of similar size as the coefficients of the neighbor effects in the control group. Indeed, we cannot reject the hypothesis that the absolute values of both coefficients are equal ($0.7907 \leq p \leq 0.8801$).

Fourth, lending credibility to our design, Table 3 suggests that the similarities in the counterfactual neighbors' answers do not differ between our experimental conditions. To see this, note that the coefficients of the non-interacted treatment indicators $\widehat{\beta}_2$ and $\widehat{\beta}_3$ in Column (1) are very small and not significantly different from zero.

Robustness Checks. The Online Appendices A and B present additional robustness checks. Table A3 controls for row effects, and Table A4 employs non-neighbors sitting in the same row as counterfactual neighbors. The results are robust. Finally, Figure B16 and Table A5 consider identical answers (correct and incorrect) as an outcome variable. All the average neighbor effects are at least as large as in our main specification (see Figure B16). However, in line with our previous results, adding identical correct answers to our outcome introduces noise to the dependent variable, resulting in broader confidence bands.

Treatment-Specific Share of Cheaters and Amount of Cheating. Following Subsection 3.4, we can also provide lower bounds for the treatment-specific share of cheaters. Figure B15 in Online Appendix B reports the corresponding treatment-wise distributions. The lower bound takes a value of 5.8% in the control group, and it equals 12.7% in the signature treatment. Under close monitoring, we do not find a significant shift of the distribution of identical incorrect answers for actual neighbors compared to that for counterfactual neighbors. This result suggests that the share of cheaters was close to zero under close monitoring. Next, we apply our back-of-the-envelope calculation to determine the amount of cheating among cheaters. The results suggest that cheaters plagiarized, on average, 8.3 wrong answers in the control group and 6.3 wrong answers in the signature treatment. Taken together, these suggestive results imply that the signature treatment converted some non-cheaters into cheaters. However, those "converted students" seem to have plagiarized, on average, less than the cheaters in the control group.

4.2 Suggestive Evidence on Channels

The finding that close monitoring eliminates cheating is in line with the aforecited literature on deterrence and Becker's (1968) seminal theory on crime and punishment. By contrast, the standard theories (including the one of Becker) have more difficulties explaining why the honesty declaration backfired. Against this backdrop, we provide suggestive evidence on the channels through which the signature treatment might have elevated cheating. The evidence comes from a follow-up experiment conducted in the years after the experiment described in Section 4 (now labeled initial experiment). For brevity, this subsection only offers a brief overview of the design and the main results. Online Appendix E presents the details.

Possible Channels. We consider three channels through which the request to sign an honesty declaration could have increased cheating. First, the request may have decreased the perceived sanctions for plagiarism. One situation in which this effect could have occurred is when students overestimated the sanction without such a request. The demand to sign the declaration may then have directed the examinees' attention to the sanction's true (lower) level. Second, the request also could have signaled that the monitoring of examinees is ineffective, lowering the students' perceived detection probability. Third, the declaration might have shifted the examinees' beliefs about their peers' honesty. In particular, they may have interpreted the declaration as a signal that cheating in exams is widespread, weakening the perceived descriptive norm of academic integrity.[32] Students with a preference to conform to their descriptive norm (Bernheim, 1994) should then also have cheated more (e.g., because they found cheating more acceptable if it is widespread).

Study Design. We conducted the follow-up experiment with two new cohorts of freshmen who took the first exam (principles of economics) in two semesters after the initial experiment. The basic idea of our design was to study how an honesty declaration that the students signed before an exam affected the students' self-reported perceptions of cheating-related sanctions, detection probabilities, and descriptive norms. More specifically, our design consisted of two elements. First, similar to our initial experiment, we induced random variation in whether or not the examinees had to sign the honesty declaration before the exam. The probability for assignment to the signature treatment was 50%. Second, we elicited the students' perceptions via a survey. To this end, two hours after the exam, we invited the examinees to participate in a paid (€3.50) five-minute online survey on "how students generally perceive exams at the university." To prevent the students from foreseeing our goal to study the impact of the honesty declaration, we did not refer to the previous exam at any point during the survey. Furthermore, rather than inviting examinees to our study through the courses' usual communication channels, we recruited them through an official mailing list that researchers at the department frequently use to invite students to complete surveys.

[32] The literature frequently highlights two types of norms (Lapinski and Rimal, 2005). Injunctive norms reflect people's perceptions about what should be done. Descriptive norms refer to beliefs about what is actually done by others.

Sample. In sum, the two new cohorts of freshmen consisted of 1060 students. Before the follow-up experiments, however, only 233 of these students signed up for the mailing list to participate in academic studies. Hence, we could only contact this subset of students for our survey. Ultimately, 103 students completed the survey within two days after the invitation. The signature treatment and the control group were balanced in observable characteristics (see Table E2 in Online Appendix E). Furthermore, the survey participants had similar characteristics as non-participants (see Table E3).

Perceived Sanctions: Results. The post-exam survey measured the students' perceived sanctions for cheating. Particularly, we asked the survey participants to indicate their expected sanction (out of a list of five options) for two hypothetical peers: one who plagiarized in their last exam and one who used unauthorized materials in their last exam. We do not find any effect of the signature treatment on the students' perceived sanctions (see Table E4 in Online Appendix E). First, independent of the treatment, a vast majority of participants correctly indicated that cheaters would have failed the last exam (plagiarism: 71.8%; unauthorized materials: 84.5%). Second, for both forms of cheating, we cannot reject the null hypothesis that the signature treatment did not affect the distributions of the perceived sanctions.

Perceived Detection Probability: Results. Our survey also investigated impacts on the students' perceived detection probabilities. To that end, we elicited students' beliefs about how many out of 100 students who (hypothetically) cheated in their last exam would have been caught. Our survey included two versions of this question: one that focused on plagiarism and one that dealt with unauthorized materials. Again, we do not find any significant impacts of the signature treatment on the students' post-exam-survey answers (see Table E5 in Online Appendix E).

Descriptive Norms: Results. In sharp contrast to the previous findings, the treated students expected their peers to have cheated more (see Table E5 in Online Appendix E). We asked the survey participants to think about their last exam and to state how many of 100 students they believed had cheated. Our results suggest that, compared to control-group students, examinees who signed the honesty declaration believed that four to five additional peers (out of 100) plagiarized or used unauthorized materials. We also asked the participants how many peers (out of 100) they believed would have cheated in hypothetical scenarios in which the detection probability would have been zero.[33] The

[33] The reason why we added these additional questions is that the perceived frequency of cheating in the exam might reflect, to some extent, the perceived sanction instead of the underlying descriptive norm. Given the previously reported results on the expected sanction, this is rather unlikely. However, because

effects become more pronounced: compared to the control group, the treated examinees stated that they believed that 15 additional students would have plagiarized and about 20 additional students would have used unauthorized materials.[34] In summary, although the survey cannot ultimately identify the mediating mechanisms, the evidence suggests that the honesty declaration has weakened the students' descriptive norms of academic integrity.

5 Conclusion

Academic cheating is a wasteful illicit activity. It is, however, difficult to measure and even harder to fight. This paper studies how to detect, document, and prevent plagiarism in exams. We offer three contributions to the literature. First, we propose approaches to identify and quantify plagiarism that rely on comparing the similarity in the answers of actual seat neighbors (who could plagiarize from each other) and non-neighbors (who could not copy from each other). Second, applying our methods to undergraduate exams, we comprehensively document plagiarism among freshmen. We find that at least 7.7% of the row-wise pairs of seat neighbors plagiarized from each other. Back-of-the-envelope calculations additionally suggest that, on average, cheating pairs increased the number of shared answers by at most 45.6%. Moreover, most of the cheating happened in pairs in which at least one student performed worse in high school than the median student. Third, by exploiting a field experiment, we demonstrate that close monitoring eliminated cheating. By contrast, requesting students to sign an honesty declaration backfired: students in the signature treatment plagiarized twice as much as students in the control group. A second round of experiments suggests that this unintended effect arises because the honesty declaration weakened the perceived social norm of academic integrity.

In conclusion, our paper demonstrates that plagiarism in exams is widespread, that methods exist to measure and quantify cheating, that close monitoring is an effective tool to foster academic integrity, and that requesting students to sign honesty declarations is not. Hence, our paper not only contributes to the academic discussion on how to identify academic cheating but also speaks to educators around the world on how to enforce (or not enforce) academic honesty. For the bigger picture, we would like to point out that similar methods might be applied to detect non-compliant behavior in other contexts. For example, researchers could use related methods to detect spillovers of enforcement treatments in networks. While our methods could be helpful beyond the academic context,

this result was unknown when designing the experiment, we responded to this measurement concern by including questions in our survey that set the expected sanction to zero.

[34] Our design of the follow-up experiment ensures that students in the signature and control group experienced the same cheating level (see Appendix E). Hence, we can rule out that the students' perceptions in the signature treatment differed because they experienced more cheating than the control-group students.

they aim at identifying forms of non-compliant behavior that generate correlations in the individuals' outcomes. For example, in our setting, we focus on plagiarism and, hence, likely underestimate the total incidence of academic cheating. Indeed, the relevance of the other types of cheating might have increased over time, especially in the pandemic years. Thus, we are looking forward to future studies examining the other forms of academic misconduct as well.

References

BECKER, G. S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy*, **76** (2), 169–217.

BEHAVIOURAL INSIGHTS TEAM (2012). *Applying Behavioural Insights to Reduce Fraud, Error and Debt*. Tech. rep., Cabinet Office, London.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B*, **57** (1), 289–300.

BERNHEIM, D. (1994). A Theory of Conformity. *Journal of Political Economy*, **102** (5), 841–877.

BLUME, L., BROCK, W., DURLAUF, S. and IOANNIDES, Y. (2011). Identification of Social Interactions. In A. B. Jess Benhabib and M. Jackson (eds.), *Handbook of Social Economics*, vol. 1, Elsevier, Amsterdam, pp. 853 – 964.

BOWERS, W. (1964). *Student Dishonesty and its Control in Colleges*. New York: Bureau of Applied Social Research, Columbia University.

CAGALA, T., GLOGOWSKY, U. and RINCKE, J. (2019). Does Commitment to a No-Cheating Rule Affect Academic Cheating?, mimeo, University of Erlangen-Nuremberg.

CAMERON, C., GELBACH, J. and MILLER, D. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics*, **90** (3), 414–427.

CARRELL, S., MALMSTROM, F. and WEST, J. (2008). Peer Effects in Academic Cheating. *Journal of Human Resources*, **43** (1), 173–207.

CHALFIN, A. and MCCRARY, J. (2017). Criminal Deterrence: A Review of the Literature. *Journal of Economic Literature*, **55** (1), 5–48.

CHEUNG, J. (2012). *The Fading Honor Code*. Tech. rep., The New York Times, April 13, 2014.

CLINGINGSMITH, D., KHWAJA, A. and KREMER, M. (2009). Estimating the Impact of the Hajj: Religion and Tolerance in Islam's Global Gathering. *Quarterly Journal of Economics*, **124** (3), 1133–1170.

DAVIS, S., GROVER, C., BECKER, A. and MCGREGOR, L. (1992). Academic Dishonesty: Prevalence, Determinants, Techniques, and Punishments. *Teaching of Psychology*, **19** (1), 16–20.

— and LUDVIGSON, W. (1995). Additional Data on Academic Dishonesty and a Proposal for Remediation. *Teaching of Psychology*, **22** (2), 119–121.

DEE, T. and JACOB, B. (2012). Rational Ignorance in Education: A Field Experiment in Student Plagiarism. *Journal of Human Resources*, **47** (2), 397–434.

DI TELLA, R. and SCHARGRODSKY, E. (2004). Do Police Reduce Crime? Estimates Using the Allocation of Police Forces After a Terrorist Attack. *American Economic Review*, **94** (1), 115–133.

DRAKE, C. A. (1941). Why Students Cheat. *The Journal of Higher Education*, **12** (8), 418–420.

DUFLO, E., GLENNERSTER, R. and KREMER, M. (2008). *Using Randomization in Development Economics Research: A Toolkit*, Elsevier, *Handbook of Development Economics*, vol. 4, chap. 61, pp. 3895–3962.

FALK, A. and ICHINO, A. (2006). Clean Evidence on Peer Effects. *Journal of Labor Economics*, **24** (1), 39–58.

FERRAZ, C. and FINAN, F. (2011). Electoral Accountability and Corruption: Evidence from the Audits of Local Governments. *American Economic Review*, **101**, 1274–1311.

FISHER, R. (1922). On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, **85** (1), 87–94.

GENEREUX, R. L. and MCLEOD, B. A. (1995). Circumstances surrounding cheating: A questionnaire study of college students. *Research in Higher Education*, **36** (6), 687–704.

GURYAN, J., KROFT, K. and NOTOWIDIGDO, M. J. (2009). Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments. *American Economic Journal: Applied Economics*, **1** (4), 34–68.

HERBST, D. and MAS, A. (2015). Peer Effects on Worker Output in the Laboratory Generalize to the Field. *Science*, **350** (6260), 545–549.

HOLLAND, P. (1996). Assessing Unusual Agreement Between the Incorrect Answers of Two Examinees Using the K-index: Statistical Theory and Empirical Support. *ETS Research Report Series*, **1996** (1), i–41.

JACOB, B. A. and LEVITT, S. D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics*, **118** (3), 843–877.

JACQUEMET, N., LUCHINI, S., ROSAZ, J. and SHOGREN, J. (2019). Truth-Telling under Oath. *Management Science*, **65** (1), 426–438.

JONES, P., BELLET, B. and MCNALLY, R. (2020). Helping or Harming? The Effect of Trigger Warnings on Individuals with Trauma Histories. *Clinical Psychological Science*, **8** (5), 905–917.

KLEVEN, H., KNUDSEN, M., KREINER, C., PEDERSEN, S. and SAEZ, E. (2011). Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark. *Econometrica*, **79** (3), 651–692.

KLING, J., LIEBMAN, J., KATZ, L. and SANBONMATSU, L. (2004). Moving to Opportunity and Tranquility: Neighborhood Effects on Adult Economic Self-Sufficiency and Health from a Randomized Housing Voucher Experiment, princeton University Working Paper No. 5.

KREMER, M. and LEVY, D. (2008). Peer Effects and Alcohol Use among College Students. *Journal of Economic Perspectives*, **22** (3), 189–206.

LAPINSKI, M. and RIMAL, R. (2005). An Explication of Social Norms. *Communication Theory*, **15** (2), 127–147.

LEVITT, S. D. (1997). Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime. *American Economic Review*, **87** (3), 270–290.

LIN, M.-J. and LEVITT, S. (2020). Catching cheating students. *Economica*, **87** (348), 885–900.

LUCIFORA, C. and TONELLO, M. (2015). Cheating and Social Interactions. Evidence from a Randomized Experiment in a National Evaluation Program. *Journal of Economic Behavior and Organization*, **115**, 45–66.

MACKINNON, J. and WEBB, M. (2017). Wild Bootstrap Inference for Wildly Different Cluster Sizes. *Journal of Applied Econometrics*, **32** (2), 233–254.

MANSKI, C. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *Review of Economic Studies*, **60** (3), 531–542.

MANSKI, C. F. (2000). Economic Analysis of Social Interactions. *Journal of Economic Perspectives*, **14** (3), 115–136.

MARTINELLI, C., PARKER, S., PÉREZ-GEA, A. and RODRIGO, R. (2018). Cheating and Incentives: Learning from a Policy Experiment. *American Economic Journal: Economic Policy*, **10** (1), 298–325.

MAZAR, N., AMIR, O. and ARIELY, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, **45**, 633–644.

MCCABE, D. (2005). Cheating Among College and University Students: A North American Perspective. *International Journal for Educational Integrity*, **1** (1).

— and TREVINO, L. (1993). Academic Dishonesty: Honor Codes and Other Contextual Influences. *Journal of Higher Education*, **64**, 522–538.

— and — (1997). Individual and contextual influences on academic dishonesty: A multicampus investigation. *Research in Higher Education*, **38** (3), 379–396.

—, — and BUTTERFIELD, K. (2001). Cheating in Academic Institutions: A Decade of Research. *Ethics & Behavior*, **11** (3), 219–232.

MIRON, A. M. and BREHM, J. W. (2006). Reactance Theory – 40 Years Later. *Zeitschrift für Sozialpsychologie*, **37** (1), 9–18.

OLKEN, B. and PANDE, R. (2012). Corruption in Developing Countries. *Annual Review of Economics*, **4** (1), 479–509.

OLKEN, B. A. (2007). Monitoring Corruption: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy*, **115**, 200–249.

PARNTHER, C. (2020). Academic Misconduct in Higher Education: A Comprehensive Review. *Journal of Higher Education Policy and Leadership Studies*, **1** (1), 25–45.

PARR, F. W. (1936). The Problem of Student Honesty. *The Journal of Higher Education*, **7** (6), 318–326.

POWER, L. (2009). University Students' Perceptions of Plagiarism. *Journal of Higher Education*, **80**, 643–662.

RAINS, S. A. (2013). The Nature of Psychological Reactance Revisited: A Meta-Analytic Review. *Human Communication Research*, **39** (1), 47–73.

ROSENBAUM, P. (2002). Covariance Adjustment in Randomized Experiments and Observational Studies. *Statistical Science*, **17** (3), 286–327.

SACERDOTE, B. (2001). Peer Effects with Random Assignment: Results for Dartmouth Roommates. *Quarterly Journal of Economics*, **116** (2), 681–704.

SCHAB, F. (1991). Schooling Without Learning: Thirty Years of Cheating in High School. *Adolescence*, **26**, 839–847.

SHU, L., MAZAR, N., GINO, F. and BAZERMAN, D., MAX SAND ARIELY (2012). Signing at the Beginning Makes Ethics Salient and Decreases Dishonest Self-Reports in Comparison to Signing at the End. *Proceedings of the National Academy of Sciences*, **109** (38), 15197–15200.

SLEMROD, J. (2019). Tax Compliance and Enforcement. *Journal of Economic Literature*, **57** (4), 904–954.

—, BLUMENTHAL, M. and CHRISTIAN, C. (2001). Taxpayer Response to an Increased Probability of Audit: Evidence from a Controlled Experiment in Minnesota. *Journal of Public Economics*, **79** (3), 455–483.

SOTARIDONA, L. and MEIJER, R. (2003). Two New Statistics to Detect Answer Copying. *Journal of Educational Measurement*, **40** (1), 53–69.

SPENCE, M. (1973). Job Market Signaling. *Quarterly Journal of Economics*, **87** (3), 355–374.

STEINDL, C., JONAS, E., SITTENTHALER, S., TRAUT-MATTAUSCH, E. and GREENBERG, J. (2015). Understanding Psychological Reactance: New Developments and Findings. *Zeitschrift für Psychologie*, **223**, 205–214.

VAN DER LINDEN, W. and SOTARIDONA, L. (2006). Detecting Answer Copying when the Regular Response Process Follows a Known Response Model. *Journal of Educational and Behavioral Statistics*, **31** (3), 283–304.

VERSCHUERE, B., MEIJER, E., JIM, A., HOOGESTEYN, K., ORTHEY, R., MCCARTHY, R. and SKOWRONSKI, J. (2018). Registered Replication Report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, **1** (3), 299–317.

WESOLOWSKY, G. (2000). Detecting Excessive Similarity in Answers on Multiple Choice Exams. *Journal of Applied Statistics*, **27** (7), 909–921.

WOLLACK, J. (1997). A Nominal Response Model Approach for Detecting Answer Copying. *Applied Psychological Measurement*, **21** (4), 307–320.

— (2003). Comparison of Answer Copying Indices with Real Data. *Journal of Educational Measurement*, **40** (3), 189–205.

— (2006). Simultaneous Use of Multiple Answer Copying Indexes to Improve Detection Rates. *Applied Measurement in Education*, **19** (4), 265–288.

Online Appendix
(not for publication)

A Tables

Table A1: Average Neighbor Effect Under Baseline Monitoring

	Dependent Variable: Indicator for Identical Answer				
	(1) Unconditional Estimates	(2) MC Controls	(3) Hall Controls	(4) Pair Controls	(5) All Controls
A: Identical Correct and Incorrect					
Actual Neighbors	0.0202	0.0202	0.0195	0.0198	0.0188
	[0.0002]	[0.0002]	[0.0008]	[0.0003]	[0.0004]
Mean for Counterfactual Neighbors			0.577		
B: Identical Incorrect					
Actual Neighbors	0.0110	0.0110	0.0112	0.0109	0.0111
	[0.0025]	[0.0025]	[0.0016]	[0.0021]	[0.0012]
Mean for Counterfactual Neighbors			0.037		
C: Identical Correct					
Actual Neighbors	0.0092	0.0092	0.0083	0.0089	0.0077
	[0.0028]	[0.0027]	[0.0107]	[0.0388]	[0.0577]
Mean for Counterfactual Neighbors			0.540		
Multiple Choice FE	No	Yes	No	No	Yes
Hall FE	No	No	Yes	No	Yes
Pair Controls	No	No	No	Yes	Yes
Number of Clusters			8		
Number of Observations			1,121,034		

Notes: This table reports estimates of the average neighbor effect on the probability that two paired students provide identical answers under baseline monitoring. Panel *A* replicates the results of the model (1) presented in Table 2. This specification regresses a dummy indicating if two students of a pair gave the same (correct or incorrect) answer to a particular multiple-choice problem on a neighbor dummy. By contrast, Panel *B* uses a dummy indicating identical incorrect answers and Panel *C* a dummy indicating identical correct answers as an outcome. The estimates rely on linear probability models. Column (1) presents the unconditional estimates. Column (2) controls for multiple-choice fixed effects. Column (3) controls for lecture-hall fixed effects. Column (4) adds two types of pair-specific variables to our baseline regression: control variables for gender combinations (a female-female dummy and a male-male dummy) and controls for high-school grade combinations (grade indicators for the better and worse student as well as interactions). Column (5) includes all the aforementioned control variables. All specifications also include an exam dummy. Moreover, the specifications define counterfactual neighbors as pairs of students in the same hall who, however, sat in different rows. Wild-cluster-bootstrap *p*-values in [brackets].

Table A2: Alternative Counterfactual: Neighbor Effect Under Baseline Monitoring

	Dependent Variable: Indicator for Identical Answer				
	(1) Unconditional Estimates	(2) MC Controls	(3) Hall Controls	(4) Pair Controls	(5) All Controls
A: Identical Correct and Incorrect					
Actual Neighbors	0.0188	0.0188	0.0187	0.0178	0.0174
	[0.0010]	[0.0010]	[0.0014]	[0.0003]	[0.0005]
Mean for Counterfactual Neighbors			0.578		
B: Identical Incorrect					
Actual Neighbors	0.0109	0.0109	0.0113	0.0099	0.0103
	[0.0027]	[0.0027]	[0.0020]	[0.0012]	[0.0009]
Mean for Counterfactual Neighbors			0.037		
C: Identical Correct					
Actual Neighbors	0.0079	0.0079	0.0074	0.0079	0.0071
	[0.0120]	[0.0119]	[0.0126]	[0.0130]	[0.0374]
Mean for Counterfactual Neighbors			0.541		
Multiple Choice FE	No	Yes	No	No	Yes
Hall FE	No	No	Yes	No	Yes
Pair Controls	No	No	No	Yes	Yes
Number of Clusters			8		
Number of Observations			140,937		

Notes: This table reports estimates of the average neighbor effect on the probability that two paired students provide identical answers under baseline monitoring. Panel *A* regresses a dummy indicating if two students of a pair gave the same (correct or incorrect) answer to a particular multiple-choice problem on a neighbor dummy. By contrast, Panel *B* uses a dummy indicating identical incorrect answers and Panel *C* a dummy indicating identical correct answers as an outcome. The estimates rely on linear probability models. Column (1) presents the unconditional estimates. Column (2) controls for multiple-choice fixed effects. Column (3) controls for lecture-hall fixed effects. Column (4) adds two types of pair-specific variables to our baseline regression: control variables for gender combinations (a female-female dummy and a male-male dummy) and controls for high-school grade combinations (grade indicators for the better and worse student as well as interactions). Column (5) includes all the aforementioned control variables. All specifications also include an exam dummy. Moreover, the specifications define counterfactual neighbors as pairs of students in the same hall *who sat in the same row*. Wild-cluster-bootstrap *p*-values in [brackets].

Table A3: Responses to the Treatments: Specifications with Row Effects

	Dependent Variable: Indicator for Identical Incorrect Answer				
	(1) Row Controls	(2) MC Controls	(3) Hall Controls	(4) Pair Controls	(5) All Controls
Signature	0.0008	0.0008		−0.0009	
	[0.8004]	[0.7999]		[0.8092]	
Monitoring	0.0022	0.0022		0.0012	
	[0.6788]	[0.6758]		[0.7183]	
Actual Neighbors	0.0073	0.0073	0.0077	0.0071	0.0075
	[0.0899]	[0.0898]	[0.0660]	[0.0840]	[0.0570]
Signature × Actual Neighbors	0.0091	0.0091	0.0085	0.0090	0.0084
	[0.0525]	[0.0516]	[0.0787]	[0.0356]	[0.0555]
Monitoring × Actual Neighbors	−0.0063	−0.0063	−0.0064	−0.0070	−0.0070
	[0.0776]	[0.0779]	[0.0527]	[0.0478]	[0.0302]
Multiple Choice FE	No	Yes	No	No	Yes
Hall FE	No	No	Yes	No	Yes
Row FE	Yes	Yes	Yes	Yes	Yes
Pair Controls	No	No	No	Yes	Yes
Mean for Counterfactual Neighbors			0.037		
Number of Clusters			11		
Number of Observations			1,412,787		

Notes: This table demonstrates how the treatments change the likelihood that two paired students provide identical incorrect answers. The estimates rely on linear probability models. Column (1) includes row indicators to control for row effects. Column (2) additionally controls for multiple-choice fixed effects. By contrast, Column (3) additionally controls for lecture-hall fixed effects. Column (4) instead adds two types of pair-specific variables to our baseline regression: control variables for gender combinations (a female-female dummy and a male-male dummy) and controls for high-school grade combinations (grade indicators for the better and worse student as well as interactions). Column (5) includes all the aforementioned control variables. All specifications also include an exam dummy. Moreover, the specifications define counterfactual neighbors as pairs of students in the same hall who, however, sat in different rows. Wild-cluster-bootstrap p-values in [brackets].

Table A4: Responses to the Treatments: Alternative Counterfactual

	Dependent Variable: Indicator for Identical Incorrect Answer				
	(1) Unconditional Estimates	(2) MC Controls	(3) Hall Controls	(4) Pair Controls	(5) All Controls
Signature	0.0008	0.0008		0.0002	
	[0.7692]	[0.7694]		[0.9566]	
Monitoring	0.0033	0.0033		0.0036	
	[0.3222]	[0.3215]		[0.4851]	
Actual Neighbors	0.0073	0.0073	0.0077	0.0067	0.0070
	[0.1372]	[0.1372]	[0.1074]	[0.0996]	[0.0848]
Signature × Actual Neighbors	0.0088	0.0088	0.0087	0.0080	0.0081
	[0.1152]	[0.1142]	[0.1248]	[0.0907]	[0.0954]
Monitoring × Actual Neighbors	−0.0095	−0.0095	−0.0099	−0.0092	−0.0101
	[0.0791]	[0.0793]	[0.0676]	[0.0379]	[0.0394]
Multiple Choice FE	No	Yes	No	No	Yes
Hall FE	No	No	Yes	No	Yes
Pair Controls	No	No	No	Yes	Yes
Mean for Counterfactual Neighbors			0.037		
Number of Clusters			11		
Number of Observations			149,776		

Notes: This table demonstrates how the treatments change the likelihood that two paired students provide identical incorrect answers. The estimates rely on linear probability models. Column (1) presents the unconditional estimates. Column (2) controls for multiple-choice fixed effects. Column (3) controls for lecture-hall fixed effects. Column (4) adds two types of pair-specific variables to our baseline regression: control variables for gender combinations (a female-female dummy and a male-male dummy) and controls for high-school grade combinations (grade indicators for the better and worse student as well as interactions). Column (5) includes all the aforementioned control variables. All specifications also include an exam dummy. Moreover, the specifications define counterfactual neighbors as pairs of students in the same hall *who sat in the same row*. Wild-cluster-bootstrap *p*-values in [brackets].

Table A5: Responses to the Treatments: All Identical Answers

	Dependent Variable: Indicator for Identical Answer				
	(1) Unconditional Estimates	(2) MC Controls	(3) Hall Controls	(4) Pair Controls	(5) All Controls
Signature	−0.0028	−0.0028		0.0003	
	[0.8658]	[0.8658]		[0.9868]	
Monitoring	−0.0448	−0.0450		−0.0443	
	[0.1548]	[0.1542]		[0.2396]	
Actual Neighbors	0.0159	0.0159	0.0140	0.0166	0.0143
	[0.0442]	[0.0442]	[0.0785]	[0.0421]	[0.0806]
Signature × Actual Neighbors	0.0105	0.0105	0.0135	0.0085	0.0115
	[0.2162]	[0.2159]	[0.1083]	[0.4070]	[0.2441]
Monitoring × Actual Neighbors	−0.0154	−0.0154	−0.0250	−0.0121	−0.0231
	[0.0710]	[0.0703]	[0.0240]	[0.1575]	[0.0545]
Multiple Choice FE	No	Yes	No	No	Yes
Hall FE	No	No	Yes	No	Yes
Pair Controls	No	No	No	Yes	Yes
Mean for Counterfactual Neighbors	0.578				
Number of Clusters	11				
Number of Observations	1,412,787				

Notes: This table demonstrates how the treatments change the likelihood that two paired students provide identical (correct or incorrect) answers. The estimates rely on linear probability models. Column (1) presents the unconditional estimates. Column (2) controls for multiple-choice fixed effects. Column (3) controls for lecture-hall fixed effects. Column (4) adds two types of pair-specific variables to our baseline regression: control variables for gender combinations (a female-female dummy and a male-male dummy) and controls for high-school grade combinations (grade indicators for the better and worse student as well as interactions). Column (5) includes all the aforementioned control variables. All specifications also include an exam dummy. Moreover, the specifications define counterfactual neighbors as pairs of students in the same hall who, however, sat in different rows. Wild-cluster-bootstrap *p*-values in [brackets].

Table A6: Monitoring Intensity by Lecture Hall

Control		Signature		Monitoring	
Hall	Students per Supervisor	Hall	Students per Supervisor	Hall	Students per Supervisor
1	51.3	*5*	56.5	*9*	9.2
2	49.8	*6*	47.5	*10*	8.5
3	38.0	*7*	44.5	*11*	8.0
4	29.0	*8*	30.0		
Treatment-specific Averages					
	46.4		46.6		8.4

Notes: This table contains information on the number of students per supervisors in each lecture hall. It also shows the weighted average of the monitoring intensity within the control group, the signature, and the monitoring treatment, respectively (weights: number of students in lecture hall).

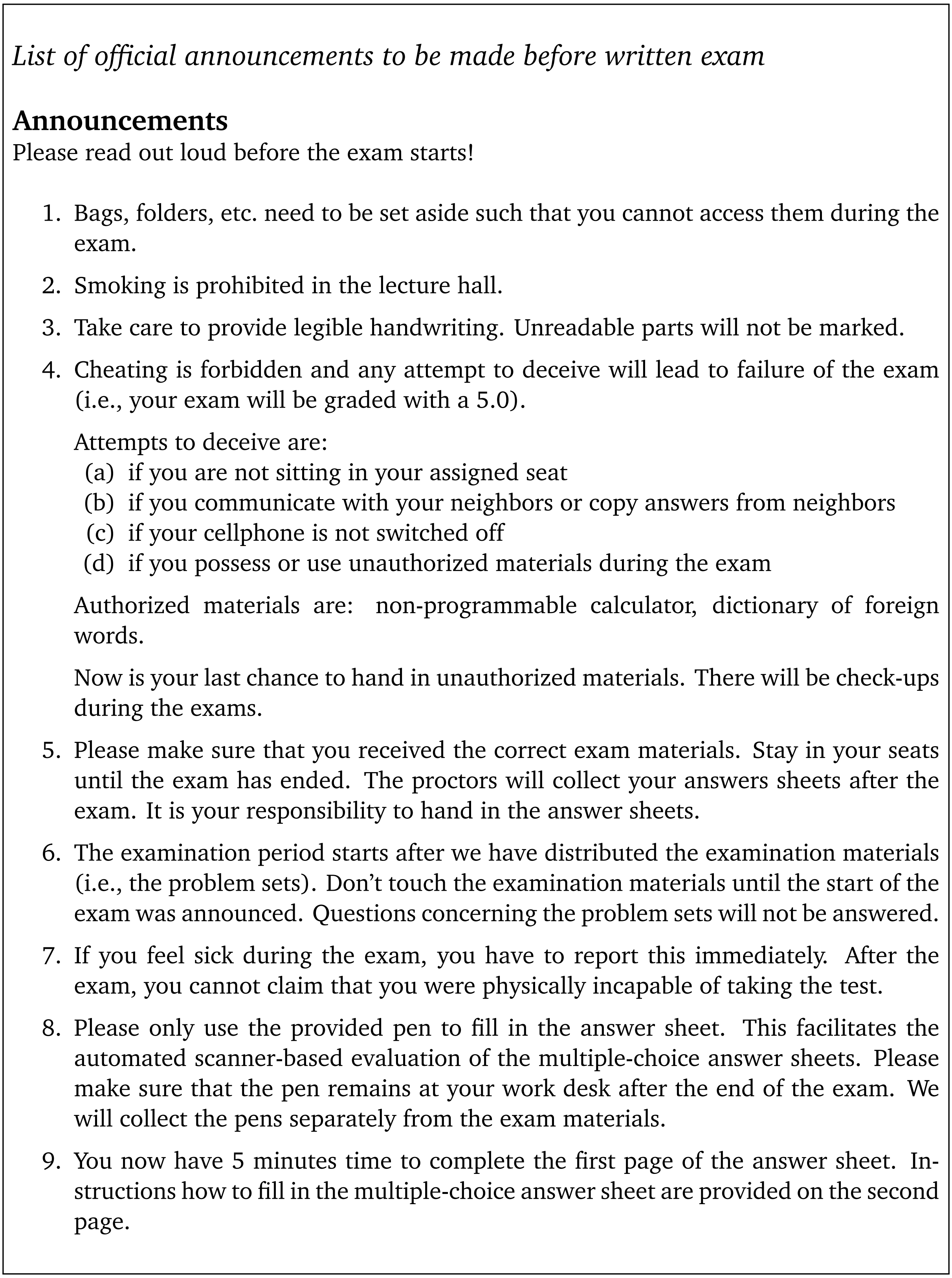
Table A7: Responses to the Signature and Monitoring Treatments in First Exam

	Dependent Variable: Indicator for Identical Incorrect Answer	
	(1) Unconditional Estimates	(2) All Controls
Signature	−0.0007 [0.8582]	
Actual Neighbors	0.0044 [0.3336]	0.0055 [0.1811]
Signature × Actual Neighbors	0.0067 [0.0957]	0.0054 [0.0889]
Multiple Choice FE	No	Yes
Hall FE	No	Yes
Pair Controls	No	Yes
Mean for Counterfactual Neighbors	0.037	
Number of Clusters	5	
Number of Observations	685,138	

Notes: This table demonstrates how the treatments change the likelihood that two paired students provide identical incorrect answers, focusing on the first exam. The estimates rely on linear probability models. Column (1) presents the unconditional estimates. Column (2) controls for multiple-choice fixed effects, hall fixed effects, and two types of pair-specific variables: control variables for gender combinations (a female-female dummy and a male-male dummy) and controls for high-school grade combinations (grade indicators for the better and worse student as well as interactions). The specifications define counterfactual neighbors as pairs of students from the same hall who did not sit next to each other. Wild-cluster-bootstrap p-values in [brackets].

B Figures

Figure B1: List of Announcements



List of official announcements to be made before written exam

Announcements

Please read out loud before the exam starts!

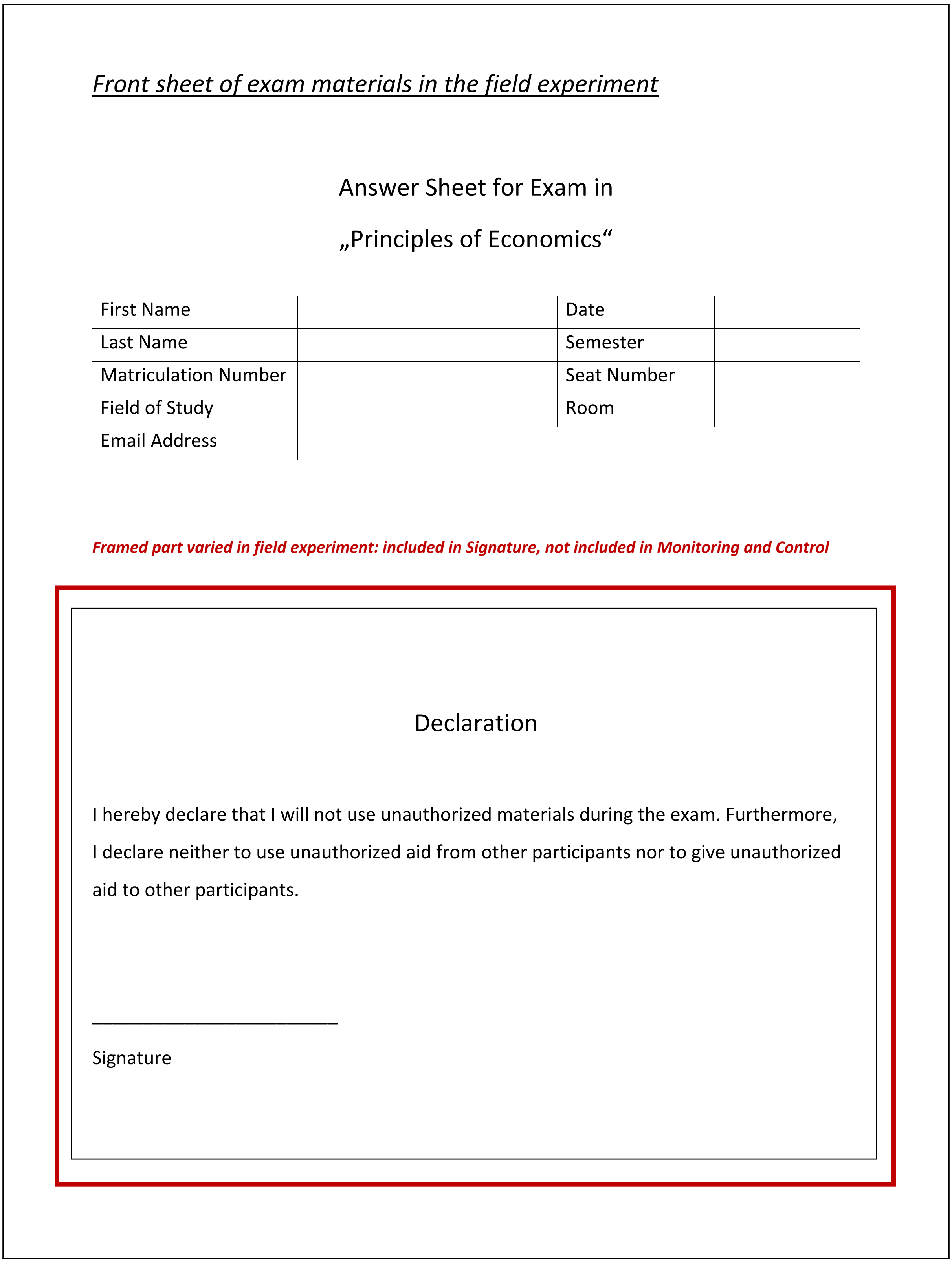
1. Bags, folders, etc. need to be set aside such that you cannot access them during the exam.
2. Smoking is prohibited in the lecture hall.
3. Take care to provide legible handwriting. Unreadable parts will not be marked.
4. Cheating is forbidden and any attempt to deceive will lead to failure of the exam (i.e., your exam will be graded with a 5.0).

   Attempts to deceive are:
   (a) if you are not sitting in your assigned seat
   (b) if you communicate with your neighbors or copy answers from neighbors
   (c) if your cellphone is not switched off
   (d) if you possess or use unauthorized materials during the exam

   Authorized materials are: non-programmable calculator, dictionary of foreign words.

   Now is your last chance to hand in unauthorized materials. There will be check-ups during the exams.
5. Please make sure that you received the correct exam materials. Stay in your seats until the exam has ended. The proctors will collect your answers sheets after the exam. It is your responsibility to hand in the answer sheets.
6. The examination period starts after we have distributed the examination materials (i.e., the problem sets). Don't touch the examination materials until the start of the exam was announced. Questions concerning the problem sets will not be answered.
7. If you feel sick during the exam, you have to report this immediately. After the exam, you cannot claim that you were physically incapable of taking the test.
8. Please only use the provided pen to fill in the answer sheet. This facilitates the automated scanner-based evaluation of the multiple-choice answer sheets. Please make sure that the pen remains at your work desk after the end of the exam. We will collect the pens separately from the exam materials.
9. You now have 5 minutes time to complete the first page of the answer sheet. Instructions how to fill in the multiple-choice answer sheet are provided on the second page.

Figure B2: Front Sheet and Honesty Declaration



Front sheet of exam materials in the field experiment

Answer Sheet for Exam in

„Principles of Economics“

First Name		Date	
Last Name		Semester	
Matriculation Number		Seat Number	
Field of Study		Room	
Email Address			

Framed part varied in field experiment: included in Signature, not included in Monitoring and Control

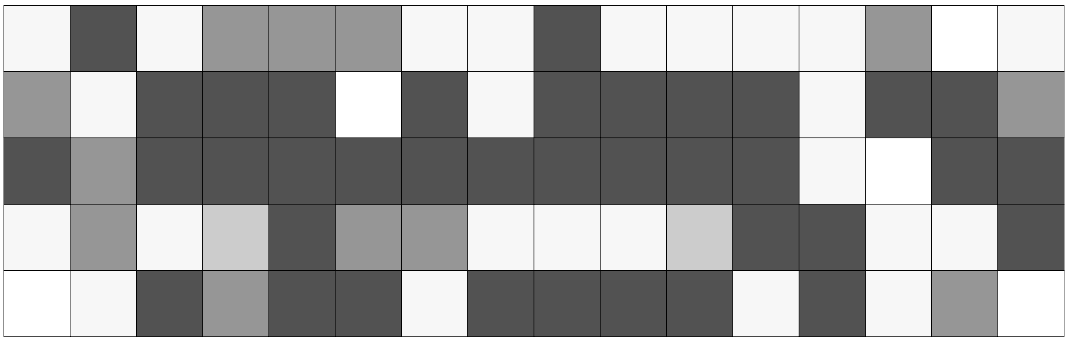
Declaration

I hereby declare that I will not use unauthorized materials during the exam. Furthermore, I declare neither to use unauthorized aid from other participants nor to give unauthorized aid to other participants.

____

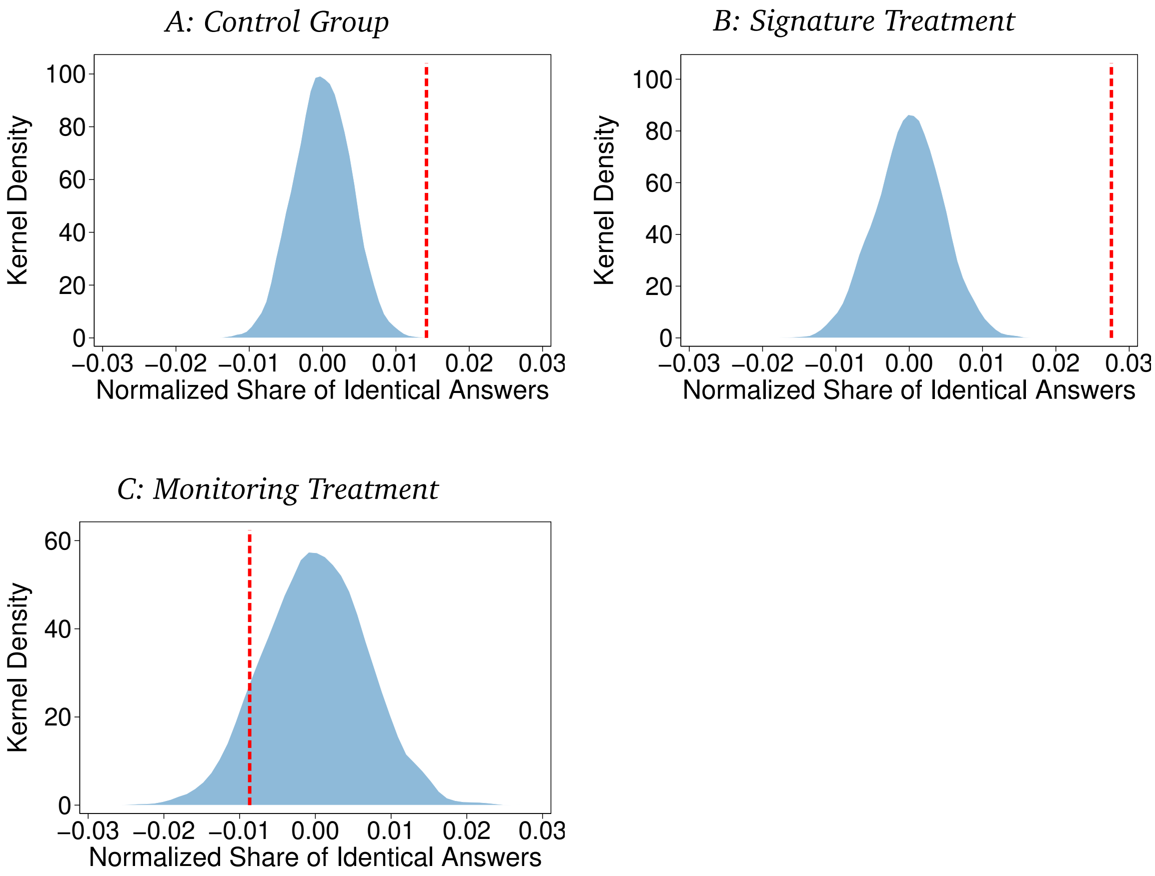
Signature

Figure B3: Responses to a Selected Multiple-Choice Problem in One Lecture Hall



Notes: This figure provides an idea of what kind of data patterns our methods exploit. It visualizes the spatial pattern of answers to one multiple-choice problem in a selected control-group hall. Each rectangle represents a student, and the shade of the rectangle indicates the student's answer. Because each multiple-choice problem consisted of four statements, there are four different shades of gray in the figure. Many students who sat next to each other provided identical answers. These correlations could reflect a spatial pattern of answers resulting from (some) students copying the responses of a direct neighbor. Such correlations could, however, also arise for other, non-cheating related reasons. For example, there could be a randomly occurring spatial pattern in the smartness of students. To evaluate whether students plagiarized, we would like to test whether the similarities in neighbors' answers were higher than in a counterfactual scenario without any cheating and only randomly occurring similarities. Our tests approximate this unobserved counterfactual by creating artificial neighbor pairs that were not sitting side by side and, thus, could not plagiarize.

Figure B4: Cheating by Treatment Group



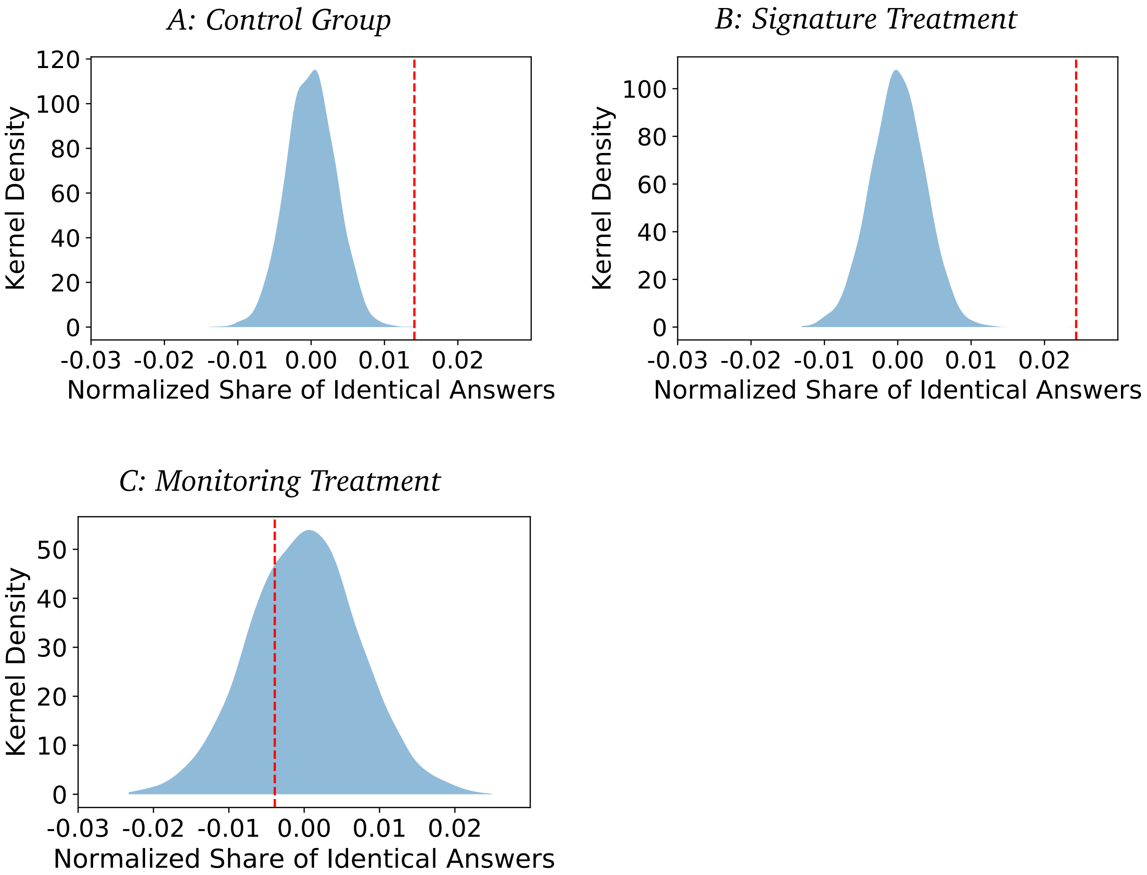
A: Control Group

B: Signature Treatment

C: Monitoring Treatment

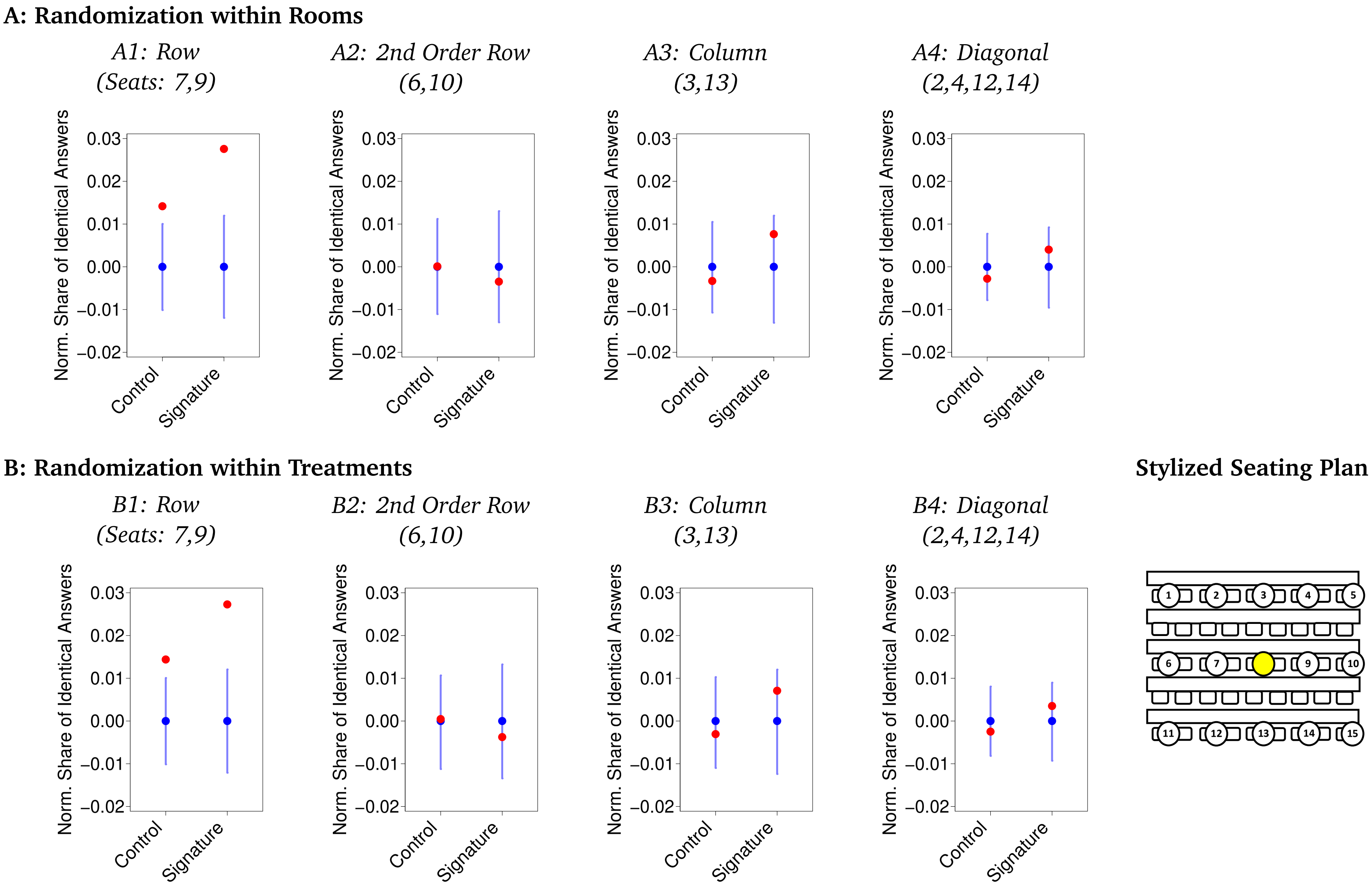
Notes: This figure shows the treatment-specific results for our randomization tests. Panel *A* focuses on the control group, Panel *B* on the signature treatment, and Panel *C* on the monitoring treatment. The vertical lines represent the test statistic derived from the actual seating arrangement. The bell-shaped curves plot the mean-centered null distribution based on Epanechnikov kernels. We obtain this distribution by randomly reassigning students within halls to seats. The reassignment procedures ensures that the counterfactual pairs do not consist of two students who were actually sitting in the same row. We obtain $p \leq 0.002$ for the control group and the signature treatment, and we find $p > 0.999$ for the monitoring treatment (two-tailed tests with Bonferroni correction).

Figure B5: Alternative Counterfactual: Cheating by Treatment Group



Notes: This figure shows the treatment-specific results for our randomization tests. Panel *A* focuses on the control group, Panel *B* on the signature treatment, and Panel *C* on the monitoring treatment. The vertical lines represent the test statistic derived from the actual seating arrangement. The bell-shaped curves plot the mean-centered null distribution based on Epanechnikov kernels. We obtain this distribution by randomly reassigning students within halls to seats. Now, counterfactual pairs can consist of two students who were actually sitting in the same row. We obtain $p \leq 0.001$ for the control group and the signature treatment, and we find $p > 0.999$ for the monitoring treatment (two-tailed tests with Bonferroni correction).

Figure B6: Spatial Structure of Cheating and Randomization Schemes



A: Randomization within Rooms

A1: Row (Seats: 7,9)

A2: 2nd Order Row (6,10)

A3: Column (3,13)

A4: Diagonal (2,4,12,14)

B: Randomization within Treatments

B1: Row (Seats: 7,9)

B2: 2nd Order Row (6,10)

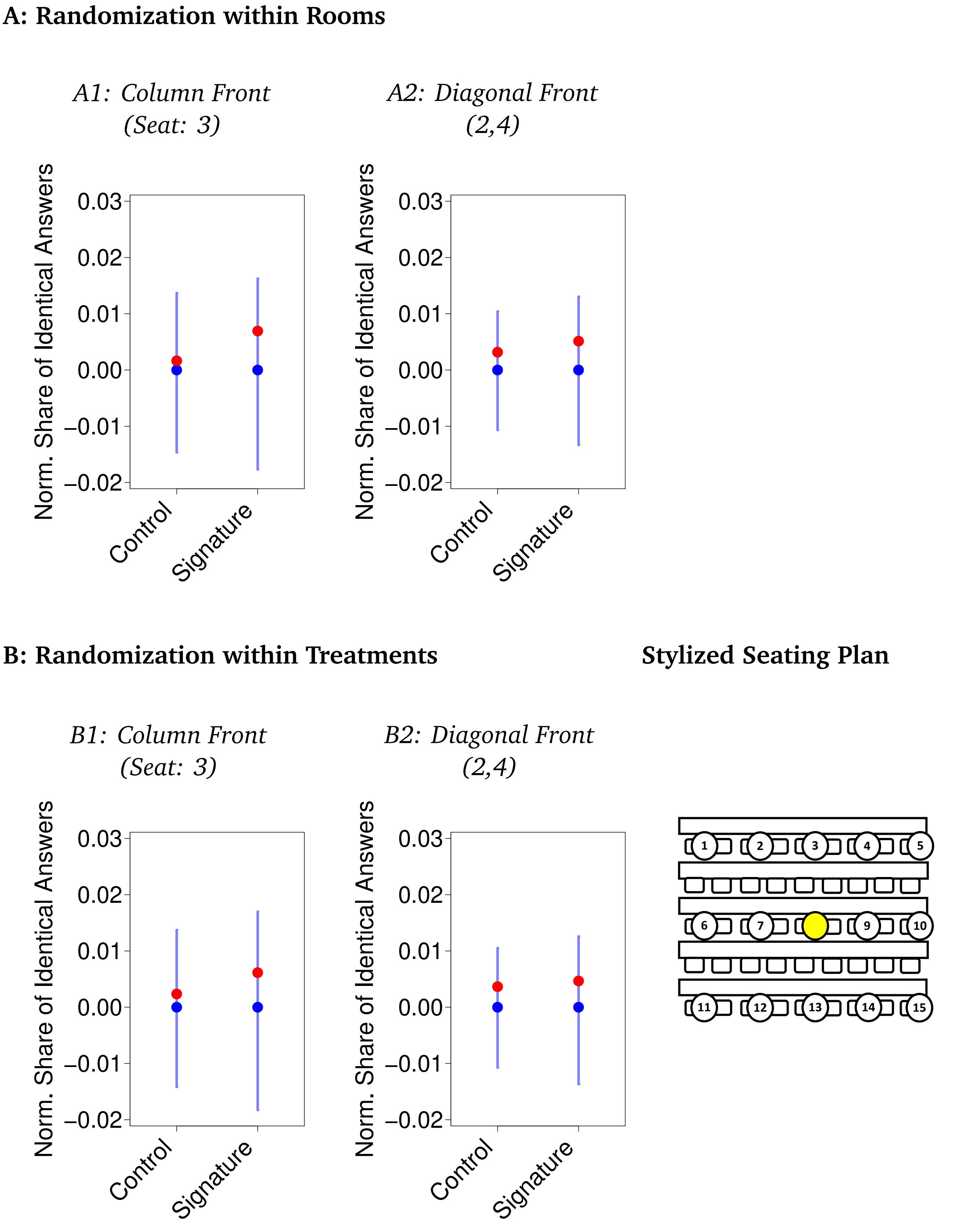
B3: Column (3,13)

B4: Diagonal (2,4,12,14)

Stylized Seating Plan

Notes: This figure serves two purposes. First, it tests the robustness of our results to the randomization schemes. Panel *A* randomizes individuals within rooms. Panel *B*, instead, resamples individuals within treatments (i.e., also across halls). Both reassignment procedures ensure that the counterfactual pairs do not consist of two students who were actually sitting in the same row. Second, the figure examines the spatial structure of cheating. The stylized seating plan helps us to highlight our specifications. The yellow circle represents a particular student (sitting in seat 8) who can copy answers from her neighbors 1 to 15. The Panels *A1* and *B1* focus on row-wise cheating of direct neighbors (student copies from 7 and 9). The Figures A2 and B2 consider plagiarizing from indirect neighbors (copying from 6 and 10). The Figures *A3* and *B3* test for column-wise cheating (copying from 3 and 13). The Figures *A4* and *B4* examine diagonal cheating (copying from 2, 4, 12, and 14). Each of the figures reports the average value of the test statistic in the counterfactual distribution after mean centering (blue circles), the 95% Bonferroni-corrected confidence bands for the counterfactual distributions (blue spikes), and the empirical value of the relevant test statistic (red circles).

Figure B7: Spatial Structure of Cheating and Randomization Schemes



A: Randomization within Rooms

A1: Column Front (Seat: 3)

A2: Diagonal Front (2,4)

B: Randomization within Treatments

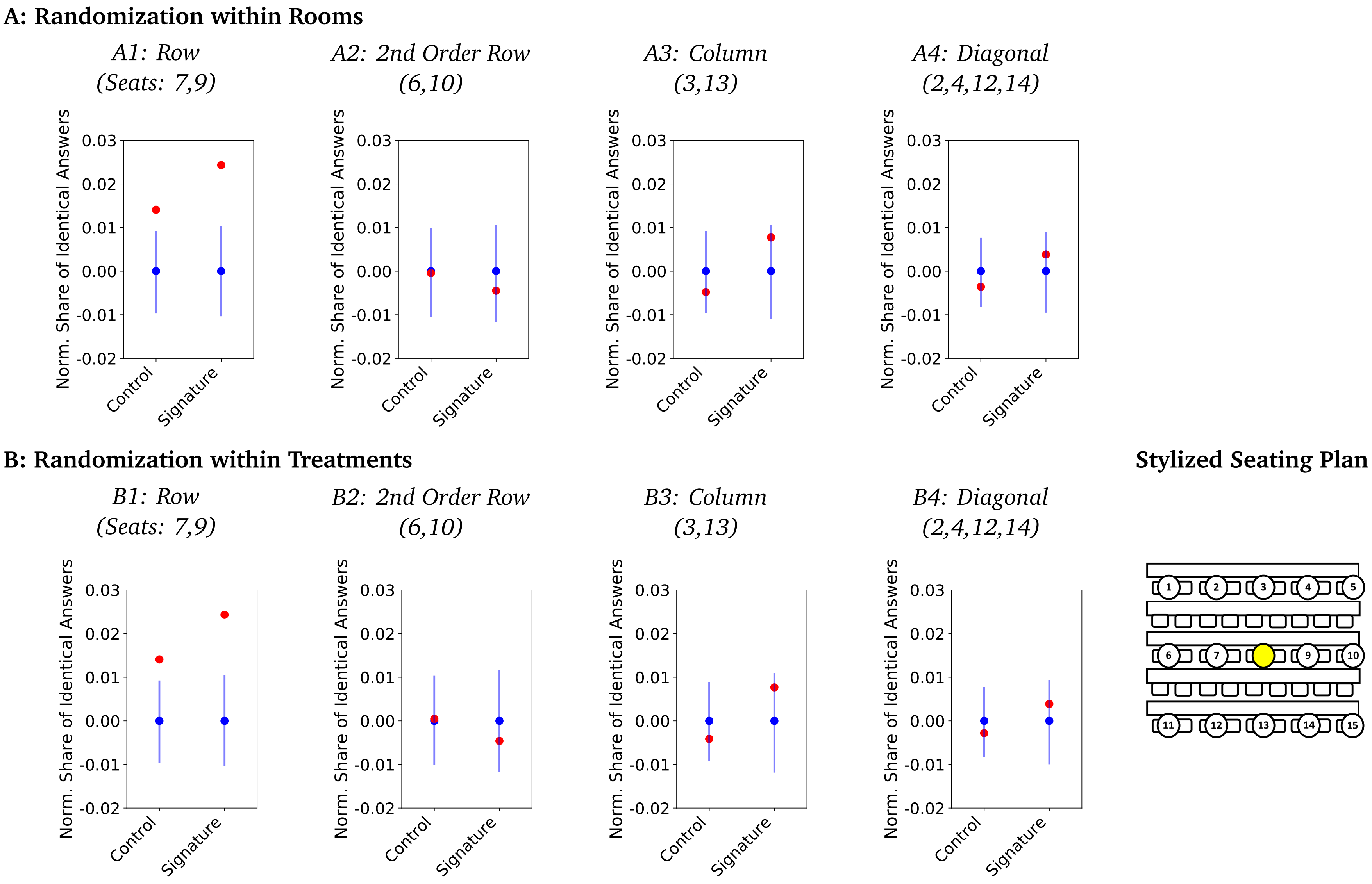
B1: Column Front (Seat: 3)

B2: Diagonal Front (2,4)

Stylized Seating Plan

Notes: This figure serves two purposes. First, it tests the robustness of our results to the randomization schemes. Panel *A* randomizes individuals within rooms. Panel *B*, instead, resamples individuals within treatments (i.e., also across halls). Both reassignment procedures ensure that the counterfactual pairs do not consist of two students who were actually sitting in the same row. Second, the figure examines the spatial structure of cheating. The stylized seating plan helps us to highlight our specifications. The yellow circle represents a particular student (sitting in seat 8) who can copy answers from her neighbors 1 to 15. The Figures *A1* and *B1* assume that the student only copied answers from the student in seat 3. The Figures *A2* and *B2* examine front-diagonal cheating (i.e., copying the answer of the students 2 and 4). Each of the figures reports the average value of the test statistic in the counterfactual distribution after mean centering (blue circles), the 95% Bonferroni-corrected confidence bands for the counterfactual distributions (blue spikes), and the empirical value of the relevant test statistic (red circles).

Figure B8: Alternative Counterfactual: Spatial Structure of Cheating and Randomization Schemes



A: Randomization within Rooms

A1: Row (Seats: 7,9)

A2: 2nd Order Row (6,10)

A3: Column (3,13)

A4: Diagonal (2,4,12,14)

B: Randomization within Treatments

B1: Row (Seats: 7,9)

B2: 2nd Order Row (6,10)

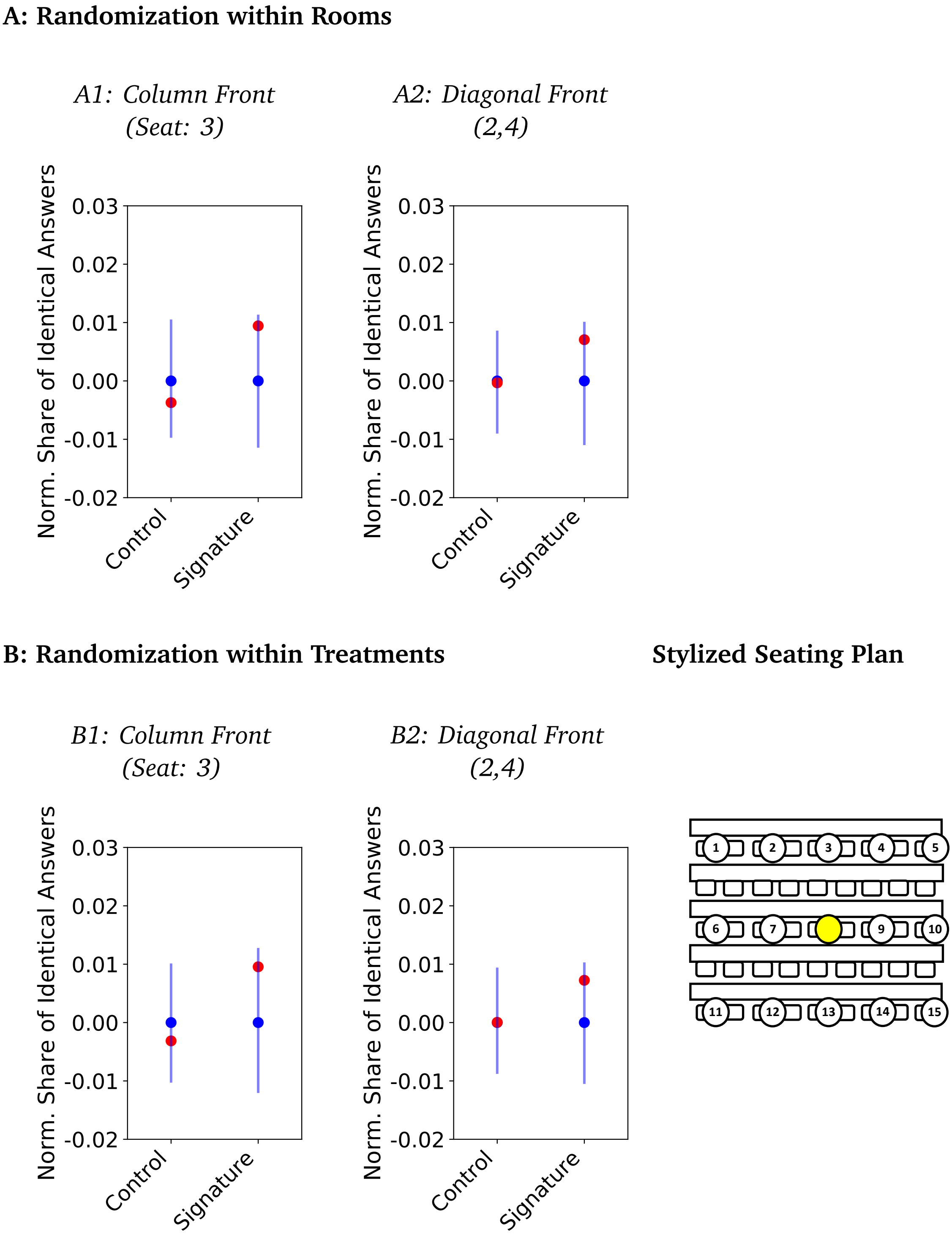
B3: Column (3,13)

B4: Diagonal (2,4,12,14)

Stylized Seating Plan

Notes: This figure serves two purposes. First, it tests the robustness of our results to the randomization schemes. Panel *A* randomizes individuals within rooms. Panel *B*, instead, resamples individuals within treatments (i.e., also across halls). Now, counterfactual pairs can consist of two students who were actually sitting in the same row. Second, the figure examines the spatial structure of cheating. The stylized seating plan helps us to highlight our specifications. The yellow circle represents a particular student (sitting in seat 8) who can copy answers from her neighbors 1 to 15. The Panels *A1* and *B1* focus on row-wise cheating of direct neighbors (student copies from 7 and 9). The Figures A2 and B2 consider plagiarizing from indirect neighbors (copying from 6 and 10). The Figures *A3* and *B3* test for column-wise cheating (copying from 3 and 13). The Figures *A4* and *B4* examine diagonal cheating (copying from 2, 4, 12, and 14). Each of the figures reports the average value of the test statistic in the counterfactual distribution after mean centering (blue circles), the 95% Bonferroni-corrected confidence bands for the counterfactual distributions (blue spikes), and the empirical value of the relevant test statistic (red circles).

Figure B9: Alternative Counterfactual: Spatial Structure and Randomization Schemes

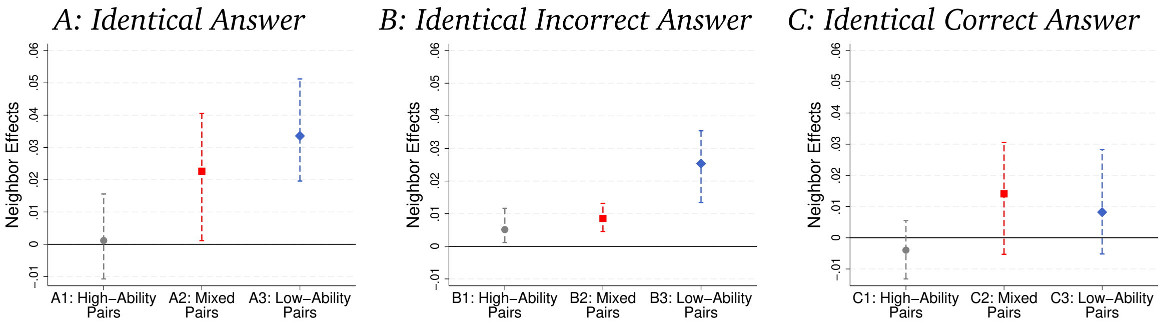

A: Randomization within Rooms

B: Randomization within Treatments

Stylized Seating Plan

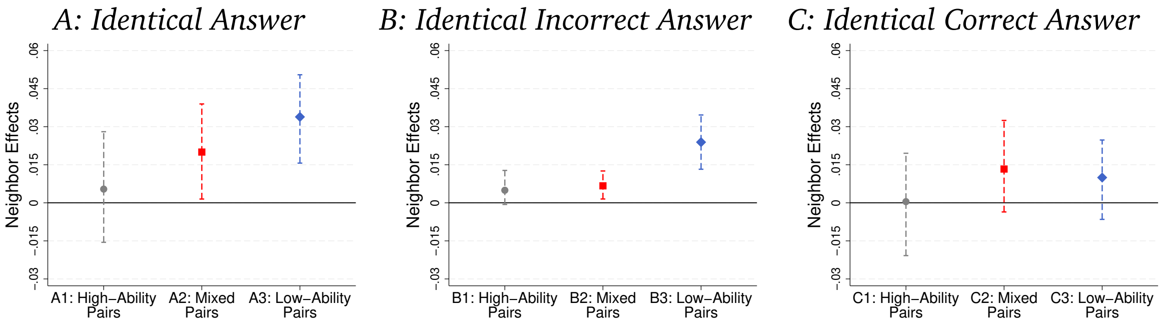
Notes: This figure serves two purposes. First, it tests the robustness of our results to the randomization schemes. Panel *A* randomizes individuals within rooms. Panel *B*, instead, resamples individuals within treatments (i.e., also across halls). Now, counterfactual pairs can consist of two students who were actually sitting in the same row. Second, the figure examines the spatial structure of cheating. The stylized seating plan helps us to highlight our specifications. The yellow circle represents a particular student (sitting in seat 8) who can copy answers from her neighbors 1 to 15. The Figures *A1* and *B1* assume that the student only copied answers from the student in seat 3. The Figures *A2* and *B2* examine front-diagonal cheating (i.e., copying the answer of the students 2 and 4). Each of the figures reports the average value of the test statistic in the counterfactual distribution after mean centering (blue circles), the 95% Bonferroni-corrected confidence bands for the counterfactual distributions (blue spikes), and the empirical value of the relevant test statistic (red circles).

Figure B10: Grade Heterogeneity: Estimation with Control Variables



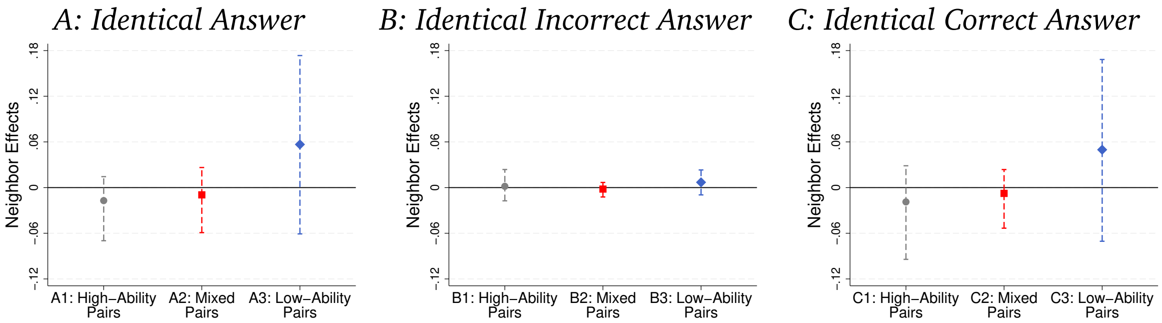
Notes: This figure examines how the students' ability (proxied by high-school performance) relates to their cheating behavior (under baseline monitoring). To construct this figure, we use model (2) to estimate the effect of being a pair of actual neighbors on the probability that two students give identical answers (Panel *A*), identical incorrect answers (Panel *B*), or identical correct answers (Panel *C*). Model (2) allows for heterogeneity in the average neighbor effect depending on a pair's ability composition. Particularly, the effects are allowed to vary in whether both students of pair *p* (gray circles), one student of pair *p* (red squares), or none of the students of pair *p* (blue diamond) performed better in high school than the "median student." The underlying regressions control for an exam dummy, multiple-choice fixed effects, and lecture-hall fixed effects. They also add two types of pair-specific variables to our baseline specification: control variables for gender combinations (a female-female dummy and a male-male dummy) and controls for high-school grade combinations (grade indicators for the better and worse student as well as interactions). All specifications derive the 95% confidence bands by a wild-cluster-bootstrap procedure.

Figure B11: Grade Heterogeneity: Definition of Grade Variables Based on Mean Student



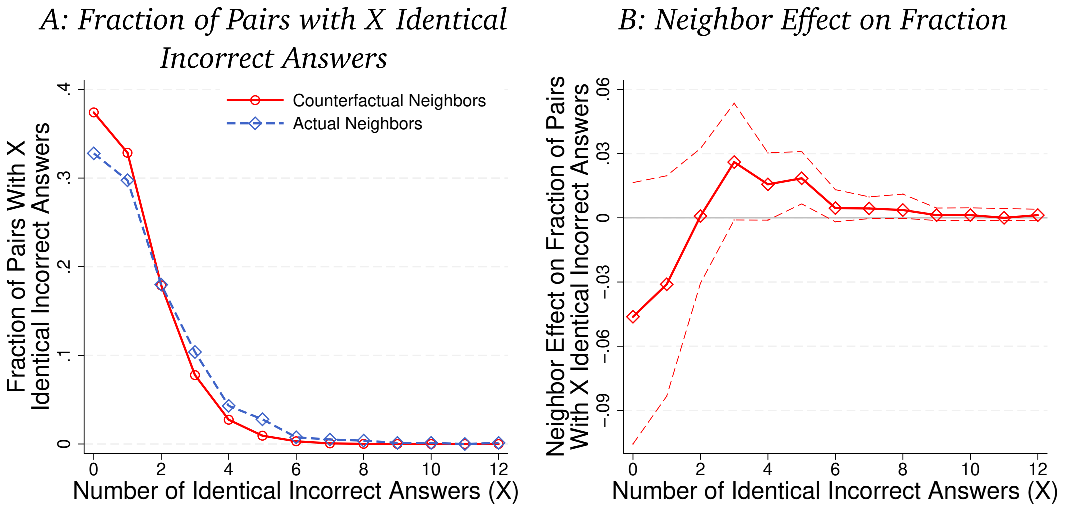
Notes: This figure examines how the students' ability (proxied by high-school performance) relates to their cheating behavior (under baseline monitoring). To construct this figure, we use model (2) to estimate the effect of being a pair of actual neighbors on the probability that two students give identical answers (Panel *A*), identical incorrect answers (Panel *B*), or identical correct answers (Panel *C*). Model (2) allows for heterogeneity in the average neighbor effect depending on a pair's ability composition. Particularly, the effects are allowed to vary in whether both students of pair *p* (gray circles), one student of pair *p* (red squares), or none of the students of pair *p* (blue diamond) performed better in high school than the "mean student." All specifications include an exam dummy and derive the 95% confidence bands by a wild-cluster-bootstrap procedure.

Figure B12: Grade Heterogeneity under Close Monitoring



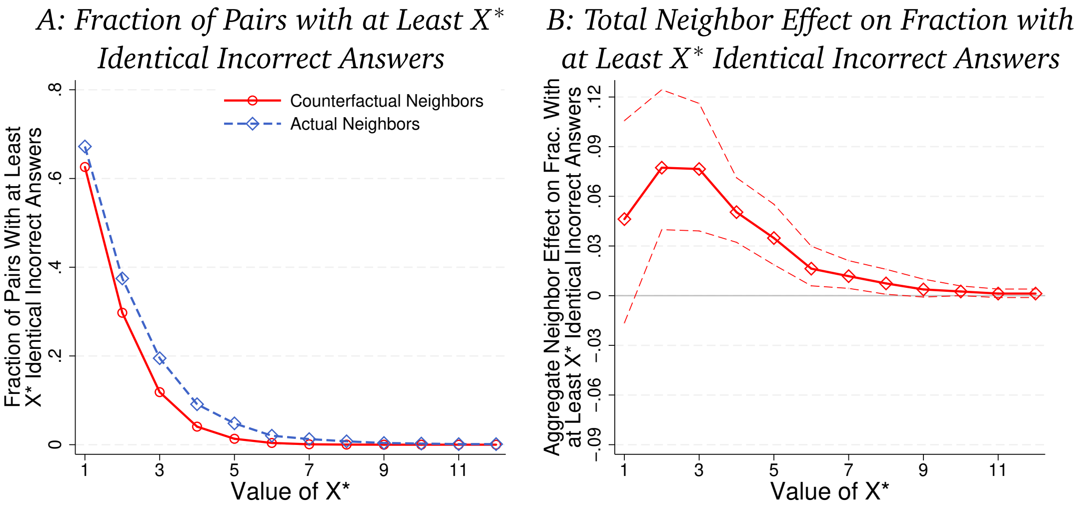
Notes: This figure examines how the students' ability (proxied by high-school performance) relates to their cheating behavior (under close monitoring). To construct this figure, we use model (2) to estimate the effect of being a pair of actual neighbors on the probability that two students give identical answers (Panel *A*), identical incorrect answers (Panel *B*), or identical correct answers (Panel *C*). Model (2) allows for heterogeneity in the average neighbor effect depending on a pair's ability composition. Particularly, the effects are allowed to vary in whether both students of pair *p* (gray circles), one student of pair *p* (red squares), or none of the students of pair *p* (blue diamond) performed better in high school than the "mean student." All specifications include an exam dummy and derive the 95% confidence bands by a wild-cluster-bootstrap procedure.

Figure B13: Shift in the Distribution: Up to Twelve Identical Incorrect Answers



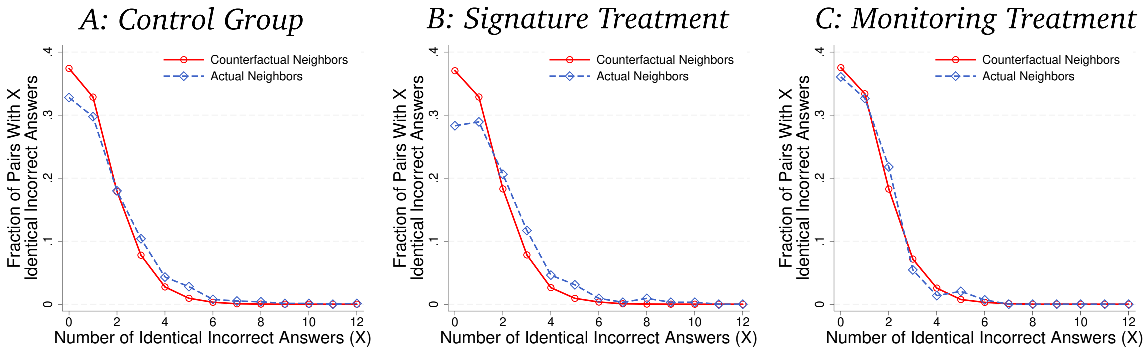
Notes: This figure shows how cheating shifts the distribution of identical incorrect answers for $X = 0$ to $X = 12$. Panel *A* depicts the distribution of identical incorrect answers for counterfactual neighbor pairs (solid red line) and actual neighbor pairs (dashed blue line). Panel *B* shows the corresponding neighbor effects, which are the X-specific differences between the actual and counterfactual distribution shown in Panel *A*. To construct the 95% confidence bands in Panel *B*, we estimate the model described in Footnote 24 and employ our standard wild-cluster-bootstrap procedure.

Figure B14: Shift in Fraction of Pairs with More Than X^* Identical Incorrect Answers



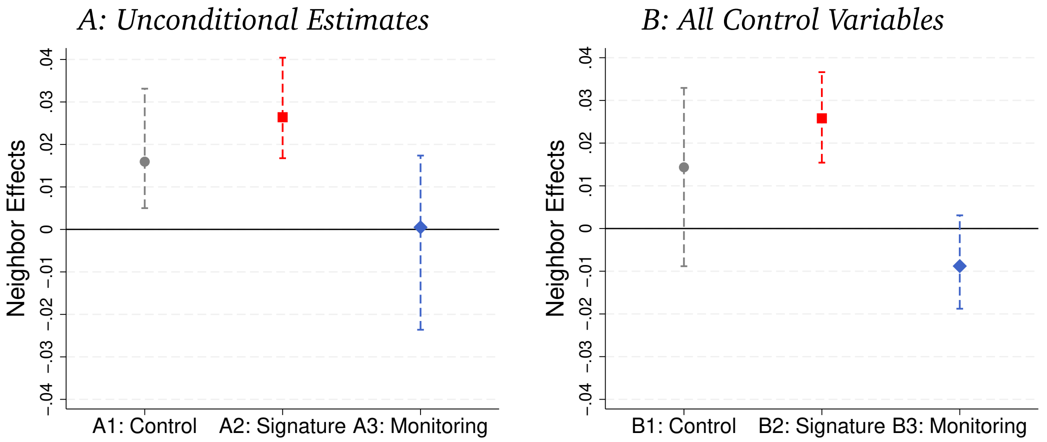
Notes: This figure shows how cheating shifts the fraction of pairs with more than X^* identical incorrect answers. The solid red line in Panel A depicts the fraction for counterfactual neighbor pairs as a function of X^*: $\sum_{X=X^*}^{30} \tilde{f}^X$. The dashed blue line shows the corresponding fraction for actual neighbor pairs: $\sum_{X=X^*}^{30} f^X$. Panel B shows the corresponding aggregate average neighbor effects, $\sum_{X=X^*}^{30} (f^X - \tilde{f}^X)$. Under our standard identifying assumptions, the aggregated neighbor effects identify a particular subset of cheating pairs: the fraction of actual neighbors that increase their number of identical incorrect answers from less than X^* answers (without cheating) to X^* or more answers (with cheating). To construct the 95% confidence bands in Panel B, we estimate an adjusted version of the model described in Footnote 24. The model regresses dummies indicating if two students of a pair p gave at least X^* identical incorrect answers on a dummy for actual neighbors. We then employ our standard wild-cluster-bootstrap procedure to derive the confidence bands.

Figure B15: Shift in the Distribution of Identical Incorrect Answers by Treatment



Notes: This figure shows how cheating shifts the distribution of identical incorrect answers. Panel *A* focuses on the control group. It depicts the distribution of identical incorrect answers for counterfactual neighbor pairs (solid red line) and actual neighbor pairs (dashed blue line). Panel *B* presents a similar graph for the signature treatment and Panel *C* focuses on the monitoring treatment.

Figure B16: Treatment Heterogeneity in the *ANE*: All Identical Answers



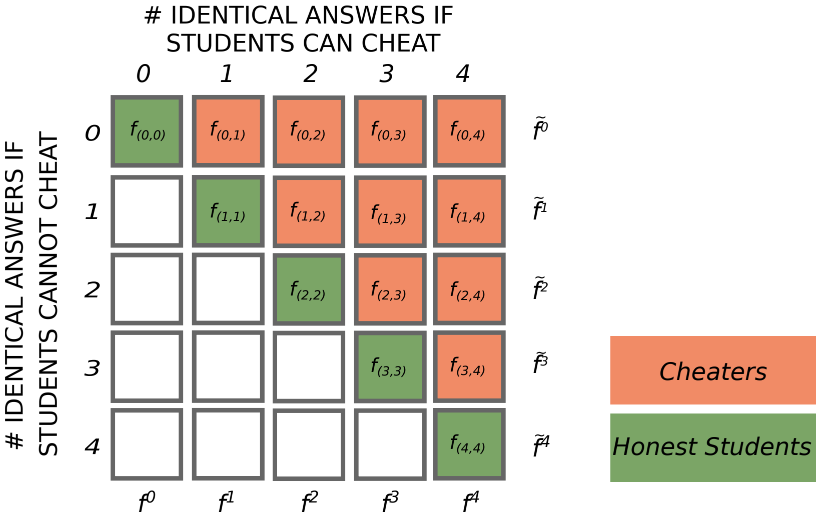
Notes: This figure shows the average neighbor effect on identical (correct and incorrect) answers for the control group (gray circles), the signature treatment (red squares), and the monitoring treatment (blue diamonds). To construct this figure, we estimate model (3) and, based on this model, predict the treatment-specific neighbor effects. Panel *A* presents the unconditional estimates. Panel *B* adds the complete set of control variables to the model. These control variables include multiple-choice fixed effects, control variables for gender combinations (a female-female dummy and a male-male dummy), and control variables for high-school grade combinations (grade indicators for the better and worse student as well as interactions). All specifications also include an exam dummy and derive the 95% confidence bands by a wild-cluster-bootstrap procedure.

C Lower Bound for the Share of Cheating Pairs

In the following, we discuss in more detail how we estimate a lower bound for the share of actual neighbor pairs that plagiarized. Our lower bound stems from comparing the distributions of identical incorrect answers for actual and counterfactual pairs.

Transition Matrix. To demonstrate why such a comparison is insightful, Figure C1 sketches a transition matrix that shows the behavior of pairs when they can and cannot cheat. For simplicity, the transition matrix considers a simplified case with only four (instead of thirty) multiple-choice problems. The rows reflect the number of identical answers in the (unobserved) counterfactual scenario (in which students cannot cheat); by contrast, the columns refer to the number of identical answers in the (observed) scenario in which the students can cheat. Moreover, the elements of the matrix represent fractions of pairs. Particularly, the value $f_{(c,p)}$ denotes the fraction of pairs that share c identical incorrect answers in the counterfactual scenario without plagiarism but p identical incorrect answers in the scenario with plagiarism. For example, $f_{(0,1)}$ is the fraction of pairs that share one incorrect answer when plagiarism is possible and otherwise zero answers.

Figure C1: Transition Matrix



# IDENTICAL ANSWERS IF STUDENTS CAN CHEAT
# IDENTICAL ANSWERS IF STUDENTS CANNOT CHEAT
Cheaters
Honest Students

	0	1	2	3	4	
0	$f_{(0,0)}$	$f_{(0,1)}$	$f_{(0,2)}$	$f_{(0,3)}$	$f_{(0,4)}$	$\tilde{f}^0$
1		$f_{(1,1)}$	$f_{(1,2)}$	$f_{(1,3)}$	$f_{(1,4)}$	$\tilde{f}^1$
2			$f_{(2,2)}$	$f_{(2,3)}$	$f_{(2,4)}$	$\tilde{f}^2$
3				$f_{(3,3)}$	$f_{(3,4)}$	$\tilde{f}^3$
4					$f_{(4,4)}$	$\tilde{f}^4$
	f^0	f^1	f^2	f^3	f^4	

Several details of the transition matrix are worth noting. First, the squares on the diagonal (marked in green) refer to pairs of students who do not cheat (they share the same number of answers in both scenarios). Therefore, the share of pairs that do not cheat corresponds to the sum of the elements on the main diagonal $f_{(0,0)}$ to $f_{(4,4)}$. Second, the above-diagonal squares (marked in red) refer to cheating pairs. For these students, the number of identical incorrect answers is higher when they can than when they cannot cheat. Consequently, the share of cheating pairs corresponds to the sum over the above-diagonal elements. Third, there are no pairs that share more identical incorrect answers when they cannot than when they can cheat (the below-diagonal elements are empty). This is true if our identifying assumption strictly holds. To see this, recap that the identifying assumption states that plagiarism is the only systematic reason why the similarity in actual and counterfactual neighbors differs.[35] By definition, plagiarism increases the similarity in the students' answers. Consequently, there are no pairs that share more incorrect answers when they cannot than when they can cheat.

Goal. Our goal is to estimate the sum over the above-diagonal elements (i.e., the share of cheaters). We propose two strategies. This Appendix presents a strategy that estimates a lower bound for the share of cheaters. To that end, we compare the distribution of identical incorrect answers for actual neighbors and the estimated distribution of counterfactual neighbors. Appendix D, instead, aims at more directly estimating the share of cheaters.

Observable Entities. Empirically, we do not directly observe all the elements of the transition matrix, which complicates the estimation for the share of cheaters. For example, we do not know how many pairs would increase their identical incorrect answers from zero to one when they can cheat. However, we can observe or estimate several entities that allow us to estimate a lower bound for the share of cheaters. First, we directly observe f^X, the fraction of actual neighbors who share X incorrect answers when they can cheat. In the transition matrix, f^X is the column sum for column X.[36] Second, using the empirical approach described in Section 3.2, we are also able to estimate $\tilde{f}^X$, the fraction of pairs sharing X incorrect answers in the scenario without cheating. This estimate approximates the row sum for row X. Third, given that we know f^X and $\tilde{f}^X$, we can estimate the neighbor effect on the distribution at value X:

$$NED^X = f^X - \tilde{f}^X. \tag{4}$$

Fourth, we are able to approximate the total aggregated neighbor effect for $X \geq X^*$:

$$TNE^{X^*} = \sum_{X=X^*} (f^X - \widetilde{f}^X). \tag{5}$$

As discussed in the following, the total aggregated neighbor effect translates into our lower bound estimate. To demonstrate why, we next discuss its interpretation.

[35] Unsystematic (random) differences would not bias our estimates.

[36] For example, the fraction of pairs that share one identical answer without cheating, f^1, consists of two types of pairs: honest students (they share one answer in both scenarios) and cheaters (they share zero answers when they cannot but one answer when they can cheat). Put differently, we have $f^1 = f_{(0,1)} + f_{(1,1)}$.

Interpretation of the Total Aggregated Neighbor Effect. The total aggregated neighbor effect, TNE^{X^*}, can be interpreted as a quantity that identifies a particular subset of cheating pairs: the share of neighbors that increase their number of identical incorrect answers from less than X^* answers (without cheating) to X^* or more answers (with cheating). This can be demonstrated mathematically. Consider, for example, the total aggregated neighbor effect TNE^1 for the simplified case with four multiple-choice problems demonstrated in Figure C1. Formally, we have:

$$TNE^1 = (f^1 - \tilde{f}^1) + (f^2 - \tilde{f}^2) + (f^3 - \tilde{f}^3) + (f^4 - \tilde{f}^4). \tag{6}$$

By plugging the definitions for f^1 to f^4 and $\tilde{f}^1$ to $\tilde{f}^4$ into equation (6)[37] and by rearranging terms, we can derive the expression:

$$TNE^1 = f_{(0,1)} + f_{(0,2)} + f_{(0,3)} + f_{(0,4)}.$$

This equation demonstrates that TNE^1 identifies the aforementioned subset of cheating pairs: it measures the share of pairs that share zero incorrect answers when they cannot plagiarize but one, two, three, or four similar incorrect answers when they can cheat (see also Figure C1). Put differently, TNE^1 corresponds to the share of neighbors that (by cheating) increase their number of identical incorrect answers from less than $X^* = 1$ answers (without cheating) to $X^* = 1$ or more answers (with cheating).

We can also calculate the total aggregated neighbor effects TNE^2, TNE^3, and TNE^4:

$$\begin{aligned} TNE^2 &= f_{(0,2)} + f_{(1,2)} + f_{(0,3)} + f_{(1,3)} + f_{(0,4)} + f_{(1,4)}, \\ TNE^3 &= f_{(0,3)} + f_{(1,3)} + f_{(2,3)} + f_{(0,4)} + f_{(1,4)} + f_{(2,4)}, \\ TNE^4 &= f_{(0,4)} + f_{(1,4)} + f_{(2,4)} + f_{(3,4)}. \end{aligned}$$

Again, the same interpretation applies. Moreover, we can generalize the argument to the more general case with thirty multiple-choice problems.

Estimating a Lower Bound. Building on the previously discussed insights, our procedure to estimate a lower bound for the share of cheaters proceeds in two steps. First, we estimate the set of total aggregated neighbor effects $TNE = \{TNE^1, TNE^2, ..., TNE^{29}\}$. The estimates rely on our identifying assumption, stating that plagiarism is the only systematic reason why the similarity in the answers of actual neighbors differs from that

[37] The definitions are (see Figure C1): $f^1 = f_{(0,1)} + f_{(1,1)}$, $f^2 = f_{(0,2)} + f_{(1,2)} + f_{(2,2)}$, $f^3 = f_{(0,3)} + f_{(1,3)} + f_{(2,3)} + f_{(3,3)}$, $f^4 = f_{(0,4)} + f_{(1,4)} + f_{(2,4)} + f_{(3,4)} + f_{(4,4)}$, $\tilde{f}^1 = f_{(1,1)} + f_{(1,2)} + f_{(1,3)} + f_{(1,4)}$, $\tilde{f}^2 = f_{(2,2)} + f_{(2,3)} + f_{(2,4)}$, $\tilde{f}^3 = f_{(3,3)} + f_{(3,4)}$, and $\tilde{f}^4 = f_{(4,4)}$.

in the answers of counterfactual neighbors.[38] Second, our lower bound for the share of cheaters corresponds the total aggregated neighbor effect TNE^{X^*} that maximizes TNE.

In a nutshell, we, hence, determine various different subset of cheating pairs: those that share more than zero identical incorrect answers with cheating, those that share more than one identical incorrect answers without cheating, and so on. The lower bound estimate for the share of cheating pairs is then given by the largest of these subsets. This strategy allows us to maximize the share of cheating pairs identifiable based on our distributional analysis.

Results. Empirically, we find that the total aggregated neighbor effect for $X^* = 2$ is the greatest. Hence, we conclude that the largest (identifiable) group of cheating pairs consists of those pairs that increase their identical incorrect answers from less than two to two or more.

Limitation. The main limitation of our distribution-based analysis is that we are only able to estimate lower bounds. The reason is that the distributions do not reveal each pair's position in the transition matrix. Thus, based on distributions, we can neither identify the elements of the transition matrix nor the share of cheaters. Instead, as discussed, the total aggregated neighbor effects only identify subsets of cheaters. Appendix D addresses this limitation. It proposes a method that allows us to identify cheating at the pair level.

[38] If this assumption holds, the estimated counterfactual distribution approximates the real counterfactual distribution in the absence of cheating. In this case, the total aggregated neighbor effects are also identified.

D Share of Cheaters: Alternative Test

Method. The distributional analysis cannot identify cheating at the pair level. However, we can alternatively measure the share of cheaters by extending the randomization tests to the pair level. The extended testing algorithm consists of four steps:

1. Calculate the test statistic as the share of all multiple-choice problems, $\widehat{s}_{i,j}$, that an examinee i and a neighbor j in the same row r answered identically.
2. Calculate the share of all multiple-choice problems $\widehat{s}_{i,m=1}$ that the examinee i shares with a counterfactual neighbor $m = 1$ who was sitting in the same hall but not in the same row.
3. Repeat the second step for all the other counterfactual neighbors $m = 2, ..., M$ who were sitting in the same hall but not the same row. This generates a distribution of $\widehat{s}_{i,m}$ with $m = 1, ..., M$ values, mean $\widehat{\mu}_{\widehat{s}}$, and standard deviation $\widehat{\sigma}_{\widehat{s}}$. This distribution corresponds to the distribution of the test statistic under the null hypothesis of no cheating by i and j.
4. Calculate the p-value as the probability that a draw from this distribution exceeds the test statistic $\widehat{s}_{i,j}$.

This approach, hence, allows to test against the null hypothesis that i does not share more identical answers with j than with the counterfactual neighbors.

Calculating the Share of Cheaters. A naive approach would then calculate the share of cheaters as the share of tests that reject the null hypothesis of no cheating at a preferred significance level (e.g., 5%). However, due to multiple testing, this strategy would overestimate the share of cheaters due to false positives. We, therefore, apply the Benjamini and Hochberg (1995) procedure such to control the false-discovery rate to the same level as the significance level (e.g., 5%).[39] For example, when setting the level to 5%, we ensure that no more than 5% of all the cases in which we reject the null hypothesis are a false discovery.

Results. We calculate the share of cheaters for multiple specifications. First, we bound the false-discovery rate to 5% and apply the testing procedure to identical (correct and incorrect) answers. Second, we restrict the rate to 1% and consider the same outcome. For both of these two specifications, we identify 8.0% of all pairs as cheaters. Third, we consider identical incorrect answers instead of all identical answers and, again, apply

[39] Note that the false-discovery rate and the false-positive rate are two different concept. The false-positive rate measures the probability of falsely rejecting the null hypothesis. Hence, it reflects the ratio between false positives to the total number of actual negative test results. By contrast, the false-discovery rate is the ratio of the number of false positives to the number of actual positive test results.

both false-discovery rates. In this case, we find a slight increase in the share of cheaters to 9.2%. This result is in line with our previous observation that identical incorrect answers contain more precise traces of cheating. We conclude that the results are fairly in line with our lower-bound estimate.

E Suggestive Evidence on Channels

This Appendix presents evidence on potential channels through which the honesty declaration might have increased cheating. Particularly, as discussed and motivated in the main body of the paper, we study if the declaration shifted the students' (a) perceived sanction, (b) perceived detection probability, and (c) perceived descriptive norms of academic integrity. However, providing evidence on these channels is particularly difficult. Not only is cheating in exams a type of behavior that is, in principle, difficult to measure, but also the individuals' perceptions are, by nature, unobservable. Besides, the institutional environment of university exams limits our options to collect data. Given these limitations (discussed in more detail below), we consider the evidence as being suggestive rather than entirely conclusive.

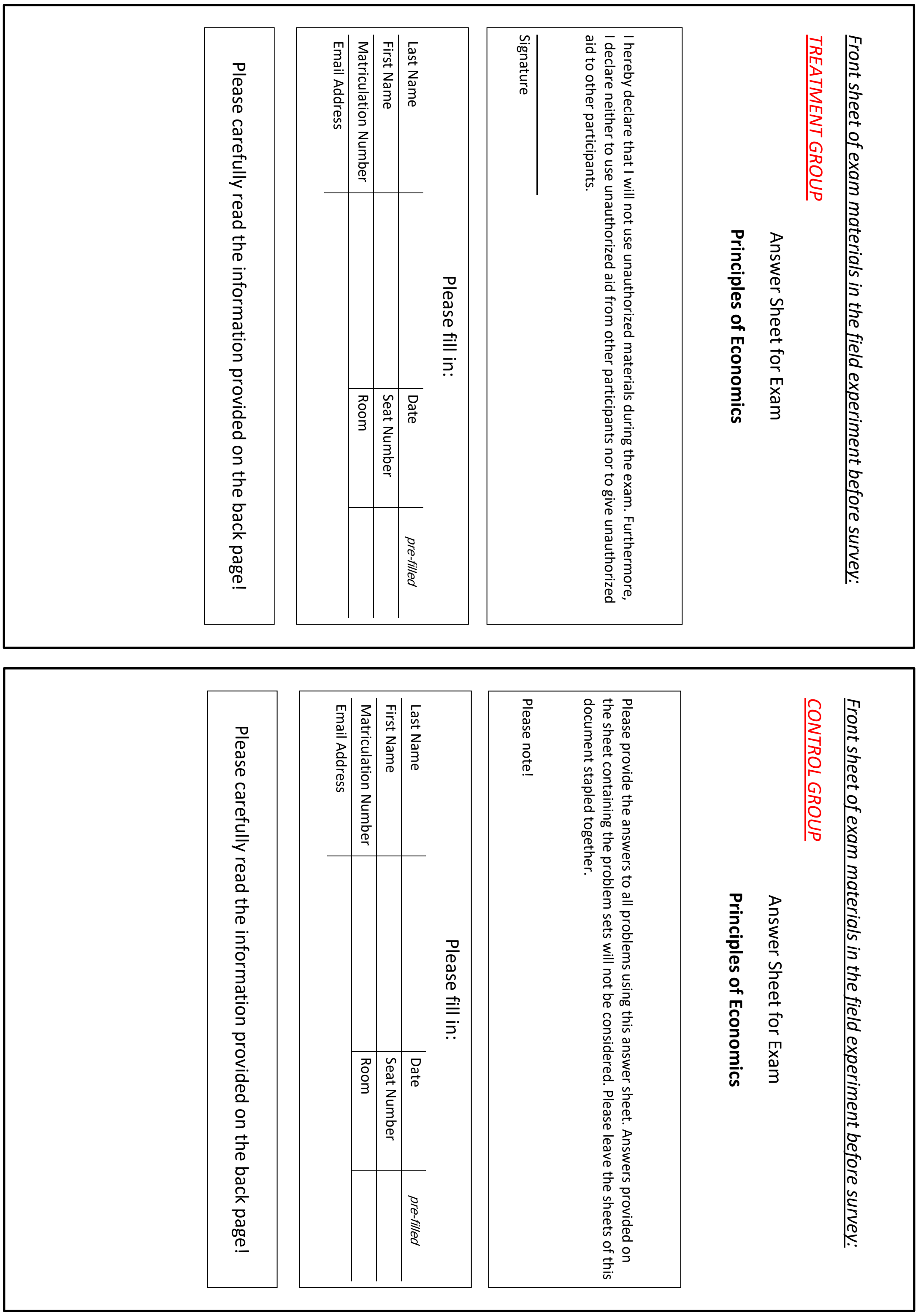
E.1 Experimental Design

We designed an entirely new follow-up experiment that allows us to collect descriptive evidence on the effects of honesty declarations on perceptions. The basic idea of this experiment was to study how an honesty declaration (that students signed before an exam) affected the students self-reported perceptions in a post-exam survey. We, hence, followed the standard approach in the literature and tackled the measurement issue of perceptions with the use of survey techniques.

Signature Treatment and Control Group: Details. We implemented our follow-up experiment in the first exam (principles of economics). As in the initial experiment, the follow-up experiment evenly split the examinees into a control group and a signature treatment. Furthermore, the signature treatment implemented an honesty declaration that was identical to the one in the initial experiment. However, in contrast to the initial experiment, we randomly assigned the treatment status within the lecture halls. Therefore, not all the students in one hall did receive the same treatment. Instead, some students were in the control group and others in the signature treatment.[40] The paragraph "Why Follow-Up Experiments?" motivates and discusses this design element in more detail.

[40] To prevent spillovers as much as possible, we kept the layout of the cover sheet of the exam materials identical between the signature treatment and the control group. Particularly, instead of the honesty declaration, the control group's cover sheet contained a text of equal length with technical information on how to handle the exam materials (see Figure E1). As a result, the exam materials looked very similar in both groups. Note that such spillovers would most likely equalize the outcomes between both conditions and, hence, would tend to downward-bias the estimated effect of the request on students' survey responses.

Figure E1: Front Sheets and Honesty Declaration: Follow-up Experiment



Front sheet of exam materials in the field experiment before survey:

TREATMENT GROUP

Answer Sheet for Exam

Principles of Economics

I hereby declare that I will not use unauthorized materials during the exam. Furthermore, I declare neither to use unauthorized aid from other participants nor to give unauthorized aid to other participants.

Signature

Please fill in:

Last Name		Date	*pre-filled*
First Name		Seat Number	
Matriculation Number		Room	
Email Address			

Please carefully read the information provided on the back page!

Front sheet of exam materials in the field experiment before survey:

CONTROL GROUP

Answer Sheet for Exam

Principles of Economics

Please provide the answers to all problems using this answer sheet. Answers provided on the sheet containing the problem sets will not be considered. Please leave the sheets of this document stapled together.

Please note!

Please fill in:

Last Name		Date	*pre-filled*
First Name		Seat Number	
Matriculation Number		Room	
Email Address			

Please carefully read the information provided on the back page!

Mailing List. We recruited students for the survey through the department's official mailing list for academic surveys. Students who sign up for this mailing list frequently receive email invitations to participate in academic surveys. Survey participants then usually get a payoff that is communicated in the invitation email and paid via bank transfer.

Survey. Two hours after the exams, we invited the examinees to participate in an online survey (via the mailing list). Students who accepted the invitation were redirected to the welcome page of the online survey. This page informed participants that the survey's goal was to measure "how students generally perceive exams at the university." It also asked participants to think about their last exam when answering the questions.[41] To prevent that students foresaw our goal to study the impact of the honesty declaration on their survey responses, we did not refer to the previous exam on principles of economics at any point during the survey. Furthermore, answering the survey took about five minutes, and participants received a flat payoff of €3.50.

Table E1: Post-Exam Survey: Questions

Perceived Sanction

S1 & S2: Imagine the supervising staff in your last exam had witnessed how one participant *copies answers from other participants* *[uses unauthorized materials (like, for instance, a smartphone)]*. What do you think would be the likely consequence for this student?

Perceived Detection Probability

D1 & D2: Think back to your last exam, and imagine 100 participants who *try to copy at least one answer from other participants* *[use unauthorized materials (like, for instance, a smartphone) to answer at least one question]*. What do you think, how many of those 100 students would have been caught? Please state a number between 0 and 100.

Descriptive Norm

N1 & N2: Think back to your last exam, and imagine a group of 100 participants. What do you think, how many of those have *copied at least one answer from other participants* *[used unauthorized materials (like, for instance, a smartphone) to answer at least one question]*? Please state a number between 0 and 100.

N3 & N4: Think back to your last exam, and imagine the supervising staff had left the exam hall for a few minutes. What do you think, how many of 100 participants would have *copied at least one answer from other participants in the meanwhile* *[used unauthorized materials (like, for instance, a smartphone) to answer at least one question in the meanwhile]*? Please state a number between 0 and 100.

Notes: This table summarizes how we measure perceptions. Each question has two versions. The first version (S1, D1, N1, N3) refers to cheating in the form of copying answers from neighbors (italics). The second version (S2, D2, N2, N4) concerns the use of unauthorized materials (gray text in brackets). To answer questions S1 and S2, participants select one of the following options: (a) There are no consequences whatsoever. (b) The student receives a verbal warning. No other consequences apply. (c) The student will face a hearing before the examination committee. (d) The committee will decide if the student fails the exam. (e) The student will fail the exam in any case. (f) The student will be relegated from the university. To answer all the other questions, participants state a number between 0 and 100.

[41] There was no other exam scheduled for freshmen students within two days after the exam in principles in economics.

Survey Questions: Details. Table E1 summarizes the survey questions. First, we elicited the perceived sanction for cheating. Particularly, students indicated their belief about the usual sanction for cheating (they choose one sanction out of a list of five). Second, we measured the perceived detection probability. To that end, we elicited beliefs about how many out of 100 cheating students would have been caught in their last exam. Third, to obtain a measure for descriptive norms, we included several questions in the questionnaire on the subjects' beliefs about the percentage of peers who cheated in the last exam. Each question came in two versions. The first version referred to cheating in the form of copying answers from neighbors and the second to the use of unauthorized materials like, for example, a mobile phone.

Table E2: Post-Exam Survey: Balancing Checks

	Control (1)	Signature (2)	Difference (3)
A: Balancing Checks for Students Who Took the Exam			
Gender (Female = 1)	0.51	0.50	0.02 (0.03)
High-School GPA	2.57	2.55	0.02 (0.04)
Math Proficiency	2.73	2.63	0.09 (0.08)
Field of Study (Econ. & Sociology = 1)	0.12	0.11	0.01 (0.02)
Age	21.2	21.4	-0.23 (0.18)
Bavaria	0.90	0.92	-0.02 (0.02)
Number of Observations	535	525	
B: Balancing Checks for Students Who Participated in the Survey			
Gender (Female = 1)	0.45	0.52	-0.07 (0.10)
High-School GPA	2.47	2.30	0.17 (0.12)
Math Proficiency	2.55	2.41	0.14 (0.24)
Field of Study (Econ. & Sociology = 1)	0.07	0.10	-0.03 (0.06)
Age	21.0	21.1	-0.06 (0.47)
Bavaria	0.89	0.90	-0.00 (0.06)
Number of Observations	55	48	

Notes: This table shows balancing checks. It reports mean values of the covariates separately for the control group (Column (1)) and the signature treatment (Column (2)). Moreover, Column (3) shows the difference in the mean values between the signature treatment and the control group with standard errors in parentheses. Panel *A* focuses on the sample of students participating in the repetition of the field experiment. Panel *B* considers the sample of individuals who participated in post-exam survey.

Sample: Details. We focused on two cohorts of students who took the exam on principles of economics in two years after our initial experiment.[42] In total, the two cohorts consisted of 1060 students of which 233 signed up for the mailing list before the exam. Ultimately, 103 students completed the survey. Panel *A* in Table E2 shows that the signature treatment and the control group are balanced in observable characteristics. Moreover, the 103 survey participants have similar observable characteristics as non-participants (see Table E3). The only significant sample imbalance is that participants have high-school GPAs which are 0.19 grade points (or 0.31 standard deviations) better than non-participants.[43] Furthermore, in the sample of survey participants, the signature treatment ($N = 48$) and the control group ($N = 55$) are well-balanced in all the observable characteristics (see Panel *B* in Table E2).

Table E3: Post-Exam Survey: Characteristics of Participants and Non-Participants

	Non-Participants	Participants	Difference
	(1)	(2)	(3)
Gender (Female = 1)	0.51	0.49	0.02 (0.05)
High-School GPA	2.58	2.39	0.19 (0.06)
Math Proficiency	2.70	2.51	0.19 (0.13)
Field of Study (Econ. & Sociology = 1)	0.12	0.09	0.03 (0.03)
Age	21.3	21.1	0.22 (0.30)
Bavaria	0.91	0.90	0.02 (0.03)
Number of Observations	957	103	

Notes: This table shows characteristics of participants and non-participants in the post-exam online survey. Column (3) shows the difference in means between non-participants and participants with standard errors in parentheses. Math proficiency is only available for 692 out of the 957 Non-participants and 83 out of 103 Participants.

Why Follow-Up Experiments? One may wonder why we needed to implement follow-up experiments to study channels. The reason is that two complications forbid us to analyze channels in our initial experiment. First, due to local exam regulations, we were not allowed to ask survey questions during the exam. After we implemented our initial experiment, the department of economics, however, established the aforementioned mailing list for academic surveys. This newly established survey allowed us to invite students who took part in the follow-up experiment to our online survey. Second, the fact

[42] The reason why we focused on two cohorts was statistical power. As in the original field experiment, we excluded students who had failed the exam previously and were not taking the exam for the first time.

[43] To probe the robustness of the survey evidence regarding this type of sample imbalance, we also estimated models that reweigh the individual observations such that the first moments of all characteristics are identical to the population estimates. The results are virtually unchanged to those reported subsequently.

that we had to survey students after the exam complicated identification. This holds especially in our initial experiment that randomized the treatments across lecture halls. To understand the potential issue of such a design, consider, for example, perceived norms as a channel. If the signature treatment increased cheating, post-exam questions on norms may reflect that students in the signature-treatment halls observe more cheating than students in control-group halls, instead of reflecting shifts in the perceived norm. This would lead us to overestimate the effect of the request on perceived norms. To tackle this issue, we adjusted the sampling scheme and, as previously mentioned, randomly assigned the signature treatment within lecture halls. This design element ensures that, on average, students in both groups experienced the same level of cheating by peers.

Drawbacks. Our design has two drawbacks. First, one obvious limitation is the limited sample size. Not all students signed up for the mailing list, and survey participation conditional on being registered is voluntary. Second, our design forbids us to reestimate the effect of the request to sign the honesty declaration on cheating. The reason is that the treatment status differs between neighbors. Similarities in their answers, thus, reflect a mix of cheating in both conditions.

E.2 Results

Perceived Sanction. We first analyze impacts on the perceived sanctions for copying answers or using unauthorized materials. Our first result is that independent of the treatment, a vast majority of students correctly indicated that cheaters would fail the exam (copying answers: 71.8%; unauthorized materials: 84.5%). Moreover, Table E4 shows no systematic differences in the perceived sanctions between the signature treatment and the control group. Using Fisher's exact tests, we cannot reject the null hypothesis that the signature treatment did not affect the distribution of answers. Hence, we do not find evidence that the honesty declaration affected perceived sanctions.

Perceived Detection Probability. Next, we study the effects of the honesty declaration on the perceived detection probability. To that end, we use the students' answers to questions *D1* and *D2* (see Table E1) as outcome variables of the model:

$$Y_{ih} = \gamma_0 + \gamma_S S_{ih} + X_{ih}\gamma_X + \pi_h + u_{ih}, \tag{7}$$

where Y_{ih} is the stated perception of student i seated in hall h, S_{ih} is an indicator for the signature treatment, X_{ih} is a vector of student controls (age, gender, and high-school GPA), and π_h absorbs exam-hall fixed effects. Regarding inference, we provide heteroscedasticity-

Table E4: Post-Exam Survey: Perceived Sanctions

	Perceived Sanction: Copying		Perceived Sanction: Unauthorized Materials	
	Control (1)	Signature (2)	Control (3)	Signature (4)
No sanction at all	0.0	0.0	0.0	0.0
Verbal warning	7.3	16.7	3.6	2.1
Exam committee hears case and decides	20.0	10.4	10.9	10.4
Student fails exam	72.7	70.8	85.5	83.3
Student is expelled from university	0.0	2.1	0.0	4.2
p-value, Fisher's exact test	[0.180]		[0.645]	
Number of Observations	55	48	55	48

Notes: This table shows students' expected sanction in case of detected cheating. Particularly, for a list of potential sanctions, the table reports the treatment-specific shares of participants (in percent) who believe that one particular sanction will be implemented in case of detection. Columns (1) and (2) focus on sanctions for copying answers. Columns (3) and (4) focus on sanctions for using unauthorized materials. We also use Fisher's exact tests to explore whether the signature treatment affected the distributions of answers. We report the corresponding *p*-values [in brackets].

consistent and wild-cluster-bootstrap *p*-values (17 clusters at the exam hall-level). In addition to the treatment effects for the individual outcomes, we also report average standardized effects according to Kling *et al.* (2004) and Clingingsmith *et al.* (2009) and exploit Mann-Whitney-U-Tests to non-parametrically test for treatment differences.

Columns (1) to (3) in Table E5 present the main results. The columns show that the signature treatment neither shifts the perceived detection probability in case of copying nor the one in case of using unauthorized materials. Taken together with the fact that the average student is well informed about the actual sanction, the absence of a significant treatment effect suggests that the signature treatment did also not significantly shift the participants' expected sanction for cheating.

Descriptive Norms To analyze the effects of the signature request on the students' descriptive norm of academic integrity, we use measures for the participants' perceived frequency of cheating (see questions *N1* and *N2*) as outcome variables in equation (7). The point estimates suggest that compared to control-group individuals, students who signed the honesty declaration believe that four to five additional peers (out of 100) plagiarized (see Column (4) in Table E5) or used unauthorized materials (see Column (5)). Only the effect of the outcome "unauthorized materials" is statistically different from zero. If we jointly exploit variation in both questions, we find a positive and significant average standardized effect regarding the perceived cheating behavior of other students.

One potential point of skepticism regarding the results on descriptive norms is that the perceived frequency of cheating in the exam may reflect, to some extent, the perceived sanction instead of the underlying descriptive norm (or the perception regarding

Table E5: Post-Exam Survey: Treatment Effects

	Detection Probability Copying	Detection Probability Unauthorized Materials	Average Stand. Effect (1) & (2)	% Other Students Copying	% Others Using Unauthorized Materials	Average Stand. Effect (4) & (5)	Social Norm Copying	Social Norm Unauthorized Materials	Average Stand. Effect (7) & (8)	Average Stand. Effect (4),(5),(7),(8)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Effect of Signature Treatment	-6.0	-2.5	-0.13	4.5	3.6	0.61	14.9	19.9	0.54	0.58
p-value, robust	[0.400]	[0.727]	[0.489]	[0.228]	[0.010]***	[0.012]**	[0.026]**	[0.004]***	[0.001]***	[0.001]***
p-value, hall cluster, wild bootstrap	[0.364]	[0.725]		[0.300]	[0.021]**		[0.052]*	[0.017]**		
p-value, Mann-Whitney-U-Test	[0.686]	[0.829]		[0.185]	[0.041]**		[0.053]*	[0.029]**		
Control Group Mean	31.6	31.4		8.6	3.4		52.2	43.4		
Number of Observations	103	103		103	103		103	103		

Notes: The table reports the effects of the signature treatment on students' responses in the post-exam survey. The estimates are derived from OLS regressions using gender, age, high-school GPA, and exam-hall fixed effects as additional controls. We report the following types of *p*-values [in brackets]: (a) heteroscedasticity robust *p*-values, (b) hall-cluster-robust *p*-values based on a wild-cluster-bootstrap procedure that accounts for the small number of cluster (Cameron *et al.*, 2008), (c) *p*-values for Mann-Whitney-U-Tests, and (d) robust *p*-values for average standardized effects following Kling *et al.* (2004) and Clingingsmith *et al.* (2009). Dependent variables in Column (1): perceived detection probability (in percent) if copying from a neighbor. Column (2): perceived detection probability (in percent) if using unauthorized materials (like smartphone, etc). Column (4): perceived share (in percent) of students copying at least one answer. Column (5): perceived share (in percent) of students using unauthorized materials. Column (7): perceived share of students (in percent) that would copy at least one answer in case of no supervision. Column (8): perceived share of students (in percent) that would use unauthorized materials in case of no supervision. See the Online Appendix for the exact wording of the survey questions.

others' perception of the sanction). Given the previously reported results on the expected sanction, this is rather unlikely. However, because this result was unknown when designing the experiment, we responded to this measurement concern by including additional questions to our survey, which introduce a hypothetical zero-enforcement scenario (see questions *N3* and *N4*). Column (7) to (9) report the results, again for both cheating technologies. Compared to Columns (4) and (5), we find a much higher level of perceived cheating in the control group, indicating that the perceived sanction, indeed, plays a role. Furthermore, for both outcomes, we confirm that the signature treatment results in a significant shift towards more (perceived) cheating by other students. The average standardized effect on perceived cheating in the zero-enforcement scenario in Column (9) is highly significant. Moreover, Column (10) displays a positive and significant average standardized effect for all four outcomes, capturing the perceived behavior of other students.

Summary. In summary, participants who signed the honesty declaration expected more cheating. In contrast to this result, we do not find any evidence for a shift in the perceived sanctions. Given the already discussed limitations that result from our inability to observe perceptions directly (limited sample size, measurement issues, spillovers), the survey evidence cannot ultimately identify the mediating mechanisms. The patterns in the data, however, suggest that the request to sign the honesty declaration has weakened the survey participants' descriptive norms of academic integrity.