Does Commitment to a No-Cheating Rule Affect Academic Cheating?

Tobias Cagala, Ulrich Glogowsky, Johannes Rincke

August 12, 2019

Abstract

Educators around the globe often require students to commit to academic integrity by signing a no-cheating declaration. This paper evaluates how such no-cheating declarations affect academic cheating. Exploiting data from a field experiment with undergraduate students, we identify cheating by comparing the similarity in multiple-choice answers of seat neighbors and counterfactual neighbors. Our main finding is that students plagiarize more after having signed a no-cheating declaration. This effect is driven by students of below-average ability. Regarding channels, we find evidence suggesting that requesting a commitment to a no-cheating rule weakens the social norm of academic integrity and triggers psychological reactance.

Keywords: academic cheating; commitment; no-cheating rule; social norm; randomization inference; field experiment

*Cagala: Deutsche Bundesbank (tobias.cagala@bundesbank.de); Glogowsky: University of Munich (ulrich.glogowsky@econ.lmu.de); Rincke: University of Erlangen-Nuremberg (johannes.rincke@fau.de). We appreciate helpful comments and suggestions by Johannes Hermle, 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.
1 Introduction

Academic cheating is a wasteful illicit activity. It distorts the incentives for students to invest in their human capital and threatens the usefulness of certificates as quality signals, undermining the efficiency of the job-matching process (Spence, 1973). Despite being harmful, academic cheating is also widespread. In surveys, between 42% and 64% of participants stated that they had cheated in college at least once (Davis and Ludvigson, 1995). Moreover, the Center for Academic Integrity at Duke University reports that, between 2002 and 2005, 21% of undergraduates admitted to having cheated on exams at least once a year (McCabe, 2005).

Given that academic cheating is harmful and prevalent, educators around the world have developed various countermeasures. One widely used instrument is asking students to commit explicitly to the norms of academic integrity. For example, when handing in assignments, term papers, or theses, students in many countries are commonly requested to sign a no-cheating declaration. Despite being commonly used, however, there is little causal evidence on the impact of commitment requests on academic honesty.

In this study, we contribute to filling this knowledge gap by providing causal evidence on how requesting a commitment to a no-cheating rule affects plagiarism on exams. The evidence originates from a field experiment carried out in written multiple-choice exams of undergraduate courses at a German university. All students in the experiment were subject to the same monitoring conditions; they also faced the same no-cheating rule that the supervisors publicly announced before the exam. As for the treatments, we randomly allocated students to a CONTROL group and a COMMITMENT group.

In the COMMITMENT treatment, before the beginning of the exam, students had to sign a declaration of compliance with the no-cheating rule. By contrast, the CONTROL group featured the university's standard exam conditions without any form of commitment.

---

1 For further evidence on the prevalence of academic cheating, see, e.g., Schab (1991) and Davis et al. (1992). Because of social-desirability biases or because subjects may not understand the principles of academic integrity, self-reported data can be biased (Power, 2009; Dee and Jacob, 2012).

2 A well-known form of commitment to academic integrity is implied by the honor code system. All of the top 10 U.S. universities according to the U.S. News & World Report 2019 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. While many practitioners question the usefulness of honor codes (Cheung, 2012), descriptive work comparing self-reported academic cheating across institutions found that cheating tends to be lower at honor-code institutions (Bowers, 1964; McCabe and Trevino, 1993; McCabe et al., 2001). To the best of our knowledge, the causal effect of honor codes on academic cheating has not been analyzed.

3 We also implemented a MONITORING treatment that imposed close monitoring of students during the exam (but no commitment). The purpose of the monitoring treatment was
Because academic cheating is a hidden activity, the first step of our analysis is to develop techniques that make cheating observable. We focus on plagiarism of multiple-choice answers, as copying solutions from seat neighbors leaves identifiable traces in the data. In particular, if students plagiarize, we expect that the similarities in seat neighbors’ answers are higher than in a counterfactual situation without any cheating and only randomly occurring similarities. In practice, the similarity in answers under the counterfactual scenario is, of course, not observable. However, we can approximate such a counterfactual scenario by considering the similarity in the answers of counterfactual neighbors (i.e., students who were not sitting side by side and, hence, could not copy from each other). Counterfactual neighbors serve as a viable control group because we randomly assigned students to seats. This feature eliminates other than cheating-related differences in the similarity of actual and counterfactual neighbors’ answers. Hence, a comparison of the similarity in the answers of actual and counterfactual neighbors identifies the amount of cheating. We exploit this basic idea in two types of tests that both simulate a large number of counterfactual neighbors: Treatment-specific non-parametric randomization tests and regression-based tests.

Our main results are as follows. First, using the treatment-specific randomization test, we document that in both the COMMITMENT and the CONTROL group, students plagiarize by copying answers from their neighbors: The similarity in neighbors’ answers is significantly higher than in the answers of counterfactual neighbors. Second, using the regression-based test, we demonstrate that cheating among low-ability students (i.e., students with poor high-school GPAs) explains the above-counterfactual similarity. Third, we exploit the regression-based test to evaluate the effect of requesting a commitment. Our results demonstrate that commitment backfires. The above-normal similarity in answers among neighbors is significantly higher in the COMMITMENT treatment than in the CONTROL group, implying that students plagiarize more in response to the commitment request.

To present suggestive evidence on the channels through which the backfiring effect of commitment operates, we repeated the experiment with a later cohort of freshmen and conducted a post-exam survey. The survey elicits the students’ perceptions of detection probabilities, sanctions, and the social norm of academic integrity. Besides studying the impact of the COMMITMENT treatment on these outcomes, we also explore whether the commitment request triggers psychological reactance, an emotional state that encourages twofold: First, under the assumption that students could cheat less under close monitoring, the treatment allows us to substantiate that our methods identify cheating. Second, in a different paper, we used the monitoring treatment to identify intertemporal spillovers of monitoring (Cagala et al., 2014). In the main part of this paper, we focus on the CONTROL group and the COMMITMENT treatment. We refer the reader to the Appendix for a full collection of results for the MONITORING treatment.
individuals to explicitly engage in forbidden activities to reassert their autonomy (Brehm, 1966). Our findings suggest that the commitment treatment weakens the perceived social norm of academic integrity. We also find suggestive evidence in line with a reactance-induced “boomerang effect”. By contrast, neither the perceived detection probability nor the expected sanction is affected by the treatment.

Our work relates to two strands of literature. First, we extend the literature studying causal effects on academic cheating. This literature shows that social interactions amplify academic cheating (Lucifora and Tonello, 2015) and that higher levels of cheating among peers lead to a higher probability that an individual cheats (Carrell et al., 2008). These findings are in line with our suggestive result that the perverse effect of requesting a commitment works through the perceived social norm of academic integrity. Other causal studies demonstrate that classroom cheating responds to monetary incentives (Jacob and Levitt, 2003; Martinelli et al., 2018). Furthermore, anti-plagiarism tutorials reduce plagiarism in term papers (Dee and Jacob, 2012), and close monitoring eliminates plagiarism in exams (Levitt and Lin, 2015).

Second, our study is linked to literature, mostly from psychology, discussing the effects of oaths, moral reminders, or commitment requests on cheating behavior. These studies typically explore the effects of such interventions on misreporting in laboratory cheating games along the lines of Fischbacher and Föllmi-Heusi (2013). Turning to the evidence, the literature suggests that primarily interventions that confront individuals with morally-charged information tend to reduce cheating (Mazar et al., 2008; Jacquemet et al., 2018; Cagala et al., 2019), although one of the main findings supporting this view has recently come under attack (Verschuere et al., 2018).

While there is, hence, some experimental work on the effectiveness of interventions that are related to commitment requests in the laboratory, there is very little evidence from the field, let alone in education-related contexts. Particularly, to the best of our knowledge, there are only two related field-experimental studies from other contexts: Shu et al. (2012) indicate that signing a no-cheating declaration 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 a no-cheating declaration from the bottom to the top of a form to apply for a tax discount may have increased fraud. Thus, it seems as if similar interventions can unfold diverging effects, depending on the context studied. In summary, the literature offers

---

4Recent contributions studying the preference for truth-telling include Kajackaite and Gneezy (2017) and Abeler et al. (2019).

5Verschuere et al. (2018) fail to replicate the finding of Mazar et al. (2008) that reminders of the Ten Commandments reduce misreporting in 19 laboratories, and, hence, question one of the most cited results suggesting that moral reminders are effective.
little guidance on whether and how to use no-cheating declarations in order to fight academic dishonesty. This observation sets the stage for our paper. Using random assignments to treatments and an objective measure of cheating, we extend both literature strands by studying how commitment requests causally impact academic dishonesty.

The structure of the paper is as follows. Section 2 analyzes the effects of commitment on academic cheating, Section 3 discusses possible channels, and Section 4 concludes.

2 Evidence on Effects of Commitment

2.1 Experimental Design

We implemented the field experiment in two written, 60-minute undergraduate exams at the business school of a German university, both of which took place in several lecture halls. The exams covered “principles of economics” (first exam) and “principles of business administration” (second exam). Both exams were compulsory for students in their first semester and were part of the curriculum for a bachelor’s degree. Because of the focus on first-year students, it is unlikely that students noticed the changes in the examination conditions that we introduced with our treatments. The department’s examination board and the lecturers who were responsible for the exams agreed to all the interventions.

As for the design of the examination questions, each exam included 30 multiple-choice problems consisting of four statements. Only one of the four statements was correct. The students’ task was to mark the correct statements on an answer sheet. All multiple-choice problems had the same weight, and the set of exam questions came in only one version. As for the design of the examination questions, each exam included 30 multiple-choice problems consisting of four statements. Only one of the four statements was correct. The students’ task was to mark the correct statements on an answer sheet. All multiple-choice problems had the same weight, and the set of exam questions came in only one version. In a given exam, every student answered the same questions appearing in the same order.

Because we are interested in dishonest behavior, in the following, we discuss general elements of the setting that might have affected the students’ decisions to cheat. A first element that was a likely driver of cheating behavior is the expected punishment in case of detection. The general rules are clear-cut: According to the exam regulations in the department, 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 supervisors in exams announce standardized examination rules by reading them aloud (see

\[6\]

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 data on student characteristics obtained from administrative records.
the Appendix for a complete list of the announcements to be made before the beginning of an exam). As part of the announcements, supervisors highlight that cheating is prohibited and that detected cheaters would fail the exam. They also emphasize a list of actions counting as cheating attempts, including copying answers from neighbors, using unauthorized materials, and not switching off mobile phones. In the experiment, we made sure that the supervisors made the announcements as planned. As a result, we believe it is justified to assume that students were similarly aware of the consequences of detected cheating in all the lecture halls. Section 3 presents evidence in line with this notion.

A second essential element affecting cheating behavior is the monitoring level, as it influences the detection probability in case of cheating attempts. Importantly, the setting we study is one in which the level of monitoring is rather low. Commonly, up to 200 students take exams in lecture halls with up to 800 seats, supervised by only two to four members of the university staff (depending on the size of the hall). Moreover, if a supervisor files a case of attempted cheating, this leads to a significant hassle during the exam and to additional paperwork with the department’s examination board after the exam. As a result, the supervising staff has little incentive to monitor students effectively. In fact, the records for the two years before the experiment show that no student failed either of the two exams because of attempted cheating.

A third element that might have affected cheating behavior (in particular, copying from neighbors) is the spatial distance between students. In the experiment, the seating arrangement was as follows: Row-wise, a student was sitting in every second seat (i.e., any two students were separated by an empty seat). Column-wise, students were sitting in every second column (i.e., any two rows with students were separated by an empty seat). The fact that the row-wise distance between two students (1.2 meters on average) was smaller than the column-wise distance (1.8 meters) or the diagonal distance (2.2 meters) suggests that students more likely copied answers from neighbors in the same row than from students sitting in the front or the back. As we demonstrate in Subsection 2.2.1, this is, in fact, precisely the spatial pattern of cheating that we find in our data.

Also of note is that the university does not have an honor code. Furthermore, in the years before the experiment, the department did not use any form of commitment requests to prevent cheating in exams.

2.1.1 Treatments

The main purpose of the field experiment is to test if commitment affects cheating in exams. To that end, we randomly allocated students from two strata (gender and high-school GPA as a proxy for ability) to one of two treatment groups: A CONTROL condition and a COMMITMENT treatment. All the students in a
given hall received the same treatment. We, thus, exclude spillovers between treatments, which substantiates the stable unit treatment value assumption. We also randomly assigned students to seats within the lecture halls and made sure that they took their preassigned seats.\footnote{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 seat, supervisors checked students IDs and made sure that the randomized seating order was put into effect.}

The only difference between the \textsc{control} group and the \textsc{commitment} treatment was that students in the \textsc{commitment} treatment signed a declaration of compliance with the no-cheating rule. We placed this declaration on the cover sheet of the exam materials (see Appendix for details). It read:

\begin{quote}
“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.”
\end{quote}

The declaration was printed below a form in which students in all treatments had to fill in their names and university IDs. The salient location was meant to direct the students’ attention to the declaration immediately before the beginning of the exam.\footnote{A post-exam check shows that all the students in the \textsc{commitment} treatment had signed the declaration of compliance.} Students were given extra time to read and complete the form and sign the declaration.

To further our understanding of the nature of commitment, two aspects of the commitment request are worth noting. First, by letting students sign the declaration, we changed the degree of commitment to an existing no-cheating rule relative to the \textsc{control} group, but neither varied the existence nor the content of the rule itself. In particular, the declaration did not introduce additional information regarding the rule. Instead, the public announcements, which were identical across treatments, laid out the rules by stating that cheating was prohibited and by highlighting the consequences of cheating. Second, the declaration was not morally loaded but neutral in the sense that it did not refer to any ethical norm.

We also implemented a treatment with close monitoring of students (but no commitment). In this paper, we use the \textsc{monitoring} treatment to substantiate that our methods can identify cheating. In particular, in the spirit of previous work, highlighting that close monitoring can eliminate academic cheating (\textit{Levitt and Lin, 2015}), we increased the monitoring intensity in the \textsc{monitoring} treatment to a level that we expected would eliminate plagiarism. In the empirical analysis, we then test whether, as expected, this type of treatment variation nullifies or, at least, sharply reduces the amount of cheating detected by our methods.
As for the implementation details of close monitoring, they were as follows: In the monitoring treatment, we allocated additional supervisors to the lecture halls such that, on average, one supervisor monitored only 8.4 students, a significant decrease relative to the 44.2 students per supervisor under baseline monitoring (in control and commitment). Importantly, in all halls, supervisors remained at specific predefined spots throughout the exam. In the control and the commitment group, supervisors took positions in the front of the hall. In the monitoring treatment, the spots where supervisors located were evenly distributed all-over the hall. Figure A1 in the Appendix provides a stylized illustration of the hall setups under baseline and close monitoring.

2.1.2 Further Details and Implementation

We took several further steps to guarantee that all examination conditions other than the treatment variations were kept constant across all the lecture halls. First, the supervising staff followed a scripted schedule, including the exact wording of all the announcements to be made before and after the exam. Second, we overbooked lecture halls when randomly allocating students to treatments, enabling us to draw students from a hall-specific pool to fill seats that otherwise would have remained empty. Due to this procedure, the actual student-per-supervisor ratios were identical to the planned ones in all the halls. 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 students before the beginning of the exam. In particular, the supervisors entered the hall and went to their preassigned positions only after all the students took their preassigned seats. As a result, on-the-spot decisions whether or not to take part in the exam should be uncorrelated with the treatment assignment. Indeed, we do not find any systematic differences in the students’ observable characteristics between treatments. Table 1 demonstrates that all characteristics are balanced across the treatments.

Next, we discuss our sampling scheme. Figure 1 presents an overview. 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 commitment 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%. In the end, 766 students took the first exam: 333 in the control group, 208 in the commitment treatment, and

---

9 Due to the overbooking procedure, some students could not be seated in their preassigned hall. We relegated those students to additional halls that we excluded from the experiment.
In the second exam, we only allocated the 432 students assigned to the first exam’s Control group to the treatments in the second exam. Hereby, we ensured that all the considered students shared a similar treatment history in the sense that they were part of the first exam’s Control group. Further note that our sample scheme aimed at maximizing the statistical power for identifying the commitment effect. We took two measures to achieve this goal. First, to increase the number of Control- and Commitment-treatment observations, we only implemented the Control and Commitment treatment in the second exam. Second, we oversampled Control-group students in the first exam, thereby increasing the total sample size (across both exams) for tests of the commitment effect. Ultimately, 353 of the 432 students from the first exam’s Control group took the second exam (204 in Control and 149 in Commitment).\(^\text{10}\)

### 2.2 Prevalence and Structure of Cheating

Like other kinds of norm-violating behaviors, cheating in exams is a hidden activity. Before we can test how commitment affects cheating behavior, we first need to develop techniques that provide evidence of a behavior that is unobservable. This section describes our approach to make cheating in exams observable. We first present a method to measure cheating and continue with applying this technique to our data, treatment by treatment. This analysis will give us a sense of whether individuals cheated at all and will inform us about the structure of cheating in our experiment. Subsection 2.3 then describes how we can also exploit the method’s underlying idea to achieve our primary goal, which is to estimate how commitment affects students’ cheating behavior.

Our identification approach of dishonest behavior in exams starts from the observation that some forms of cheating leave traces in the data that allow for inference on cheating itself. In the spotlight of this paper are traces that result from plagiarism (i.e., copying the answers of neighbors). It is important to note that 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, this will not invalidate the conclusions of our paper as long as the treatment effects are uncorrelated with the cheating technology.

Figure 2 provides an idea of what kind of data patterns our methods exploit. The figure visualizes the spatial pattern of answers to one multiple-choice problem in a selected Control-group hall. Each rectangle represents

\(^{10}\)The reason for the remaining differences in the number of students in the Control and Commitment treatment in the second exam is differences in the capacity of the lecture halls.
a student, and the shade of the rectangle indicates the student’s answer.Because each multiple-choice problem consisted of four statements, there arefour different shades of gray in the figure. It is apparent that 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. Although we randomized seats, 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 giving rise to a spatial correlation in neighbors’ answers. To evaluate whether students plagiarized, we would like to test whether the similarities in neighbors’ answers were higher than in a counterfactual situation without any cheating and only randomly occurring similarities. In practice, the counterfactual is, of course, not observable. Instead, we must find ways to approximate how the similarities would look like in the absence of cheating.

In the following, we describe our approach to tackling this problem. In particular, we propose two tests for plagiarism in exams: A non-parametric randomization test and a regression-based test. While both tests are different in nature, they share a similar core. They both compare the similarity in the answers of actual neighbors with the similarity in the answers of counterfactual neighbors (i.e., students who were not sitting next to each other). Comparing actual to counterfactual neighbors is useful for a simple reason. Counterfactual neighbors were not sitting side by side and, hence, could not copy answers from each other. This property allows us to approximate the similarity in actual neighbors’ answers in the absence of cheating, providing us with a viable control group that approximates the counterfactual situation.

To structure the discussion under which conditions the comparison of actual and counterfactual neighbors identifies plagiarism in exams, we note that the identification of cheating is closely related to identifying social effects (see, e.g., Manski 2000; Blume et al. 2011; Herbst and Mas 2015 for literature reviews). In particular, considering Manski’s (1993) conceptual considerations on the identification of social effects, it becomes clear that two types of correlations in neighbors’ answers complicate the identification of plagiarism. First, by chance or due to self-selection, neighbors may have had similar individual characteristics. For example, if students could freely choose their seats, similarly skilled individuals who tend to give similar answers might have self-selected into adjacent seats. Second, neighbors may have faced specific institutional environments during the exams, leading to stronger correlations in their answers relative to non-neighbors. An example would be day-specific examination conditions that might have aligned the students’ answers in a hall. Had we implemented a given exam on dif-

---

11In Section 2.2.1, we demonstrate that the assumption that only students sitting next to each other in the same row plagiarize from each other is supported by the data.
ferent days, comparing neighbor pairs who took the exam at one particular day to non-neighbors who took the exam at other days would invalidate the identification of plagiarism.

Having discussed the potential non-cheating induced correlations in neighbors’ answers, we can state the identifying assumption under which a comparison of actual and counterfactual neighbors identifies cheating. We have to assume that copying answers from a neighbor was the only systematic reason why the similarity in the answers of actual neighbors differed from that in the answers of counterfactual neighbors. Following up on the previous discussion on social effects, we note this is the case if the composition of both types of pairs was identical, and both types of pairs faced the same institutional environments. Put differently, all the confounding factors were equalized across the pair types. In more technical terms, our identifying assumption is that, in the absence of cheating, the similarity in a pair’s answers was independent of whether it consisted of actual or counterfactual neighbors.

We took two measures to ensure that the identifying assumption holds. First, to guarantee that there were no systematic differences in the composition of pairs, we randomly assigned individuals to seats. We, hence, followed the standard approach in the social-effects literature and exploited a randomization scheme to allocate individuals to groups within which social effects 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 construct counterfactual neighbor-pairs using only non-neighbors who, depending on the model specification, either sat in the same lecture hall or in the same row. This nets out lecture hall and row effects, respectively.

2.2.1 Cheating in Exams

In the following, we identify treatment-specific cheating behavior using a spatial randomization test. Randomization testing goes back to Fisher (1922) and is a standard inference tool in the analysis of experiments. The key characteristic of this type of test is that, instead of assuming that the test statistic follows a standard distribution under the null hypothesis, its distribution is generated from the data by resampling. In practice, randomization testing is widespread. For example, it is frequently used to derive randomization inference for treatment effect variation (Rosenbaum 2002; Duflo et al. 2008). More closely related to this paper are studies that use randomization schemes to test how outcomes of individuals are connected. For example, researchers have proposed randomization tests for specifying whether one individual’s treatment status indirectly impacts another individual’s outcome (Athey et al. 2015; Athey and Imbens 2017). Furthermore, Falk and Ichino (2006) use an
approach that is similar to ours to identify peer effects in co-workers’ productivities.

**Testing Procedure** Our paper exploits a post-experiment randomization scheme to test against the null hypothesis of no above-normal similarity in the answers of actual neighbors. In particular, we examine whether a measure for the similarity in neighbors’ answers (i.e., the test statistic) is unusually high compared to the distribution of this measure in the absence of cheating. Ex-ante, this distribution is not known. However, if the identifying assumption holds, we can approximate the distribution under no cheating by simulating a large number of artificial seat assignments (i.e., by creating counterfactual pairs of neighbors) and then recalculating what the similarity measure would have been if these assignments had been the real ones.\footnote{Our test follows the educational measurement literature in testing for cheating by examining whether the responses of two students are unusually similar (see, e.g., Holland, 1996; Wollack, 1997, 2003, 2006; Wesolowsky, 2000; Sotaridona and Meijer, 2003; van der Linden and Sotaridona, 2006). However, the standard methods in this literature test for cheating by pairs of students which educators suspect of cheating. Our methods, instead, test for plagiarism in a large population of possible pairs of students.}

Our preferred parameterization of the randomization test is one that (a) identifies plagiarism between direct neighbors who were sitting next to each other in a row and (b) randomly reassigns students within halls. We check the robustness of our results regarding these assumptions.

We now describe the details of our testing procedure. To that end, we denote by $i$ a student in row $r$ with a left neighbor $i-1$ and a right neighbor $i+1$. Our testing procedure consists of four steps:

1. Calculate the share of all multiple-choice problems $s_{i,i-1}$ that $i$ and $i-1$ answered identically (correct or incorrect). Do the same for $i$ and $i+1$ to derive $s_{i,i+1}$. Compute the treatment-specific test statistic as:

$$\hat{\Delta} = \frac{1}{N} \sum_{i=1}^{N} \frac{s_{i,i-1} + s_{i,i+1}}{2},$$

where $N$ is the number of students in the considered treatment.

2. Create counterfactual neighbor pairs by randomly reassigning students within halls to seats (without replacement), and compute the similarity measure for counterfactual neighbors, $\hat{\Delta}_{c,m=1}$.

3. Repeat the previous step $M$ times. This generates a distribution of $\hat{\Delta}_{c,m}$ with values $m=1,\ldots,M$, mean $\mu_{\hat{\Delta}_c}$, and standard deviation $\sigma_{\hat{\Delta}_c}$. Intuitively, this distribution corresponds to the distribution of the test statistic under the null hypothesis of no cheating.

4. Calculate the $p$-value of a two-tailed test as twice the probability that a draw from this distribution exceeds $\hat{\Delta}$.
**Results**  Pooling observations from both exams, we report the results of our treatment-specific randomization tests in Figure 3 ($M = 5000$). For ease of exposition, the figure reports mean-centered values (i.e., it shows $\Delta - \tilde{\mu}_\Delta$ and $\tilde{\Delta}_{c,m} - \tilde{\mu}_\Delta$). This type of normalization allows us to interpret the test statistic intuitively as the extent by which the share of identical answers among actual neighbors’ differs from the expected share for counterfactual neighbors (in percentage points).

Panel A in Figure 3 shows the results for the **CONTROL** group, and Panel B focuses on the **COMMITMENT** treatment. In each panel, 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.

Two findings emerge from Figure 3. First, Panel A shows that in the **CONTROL** group, the similarity in the answers of actual neighbors is excessively high compared to the counterfactual distribution. The test statistic indicates that the share in actual neighbors’ identical answers is about 1.4 percentage points higher than the expected share for counterfactual neighbors (who cannot cheat). This value is located in the far right tail of the counterfactual distribution, and we can, consequently, clearly reject the null hypothesis of no above-normal similarity in the answers of actual neighbors ($p$-value $= 0.001$). This finding is the first piece of direct evidence that in the **CONTROL** group, students copied answers from students sitting next to them in the same row.

Second, and most importantly, Panel B shows that we can also reject the null hypothesis of no cheating in the **COMMITMENT** treatment ($p$-value $< 0.001$). The value of the test statistic reveals that the average share of identical answers given by neighbors is about 2.4 percentage points higher than the expected share for counterfactual neighbors. We will discuss the effect of commitment at length in Subsection 2.3, but we can already highlight the intermediate result that, in our exams, the commitment request did not eliminate cheating.

To substantiate that our randomization tests identify plagiarism, we also study spatial correlations in the **MONITORING** treatment. The idea of this validation check is simple: It is natural to assume that close monitoring reduced or even eliminated the students’ options to copy answers from neighbors. Hence, if our randomization test successfully identifies plagiarism, we would expect no above-normal spatial correlations in the **MONITORING** treatment. The latter is precisely what we find: As documented in Figure A2 in the Appendix, the test statistic for the **MONITORING** treatment is located in the center of the counterfactual distribution, and we cannot reject the null hypothesis.

---

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.
hypothesis of no above-normal similarity in the answers of actual neighbors (p-value 0.999). In sum, Figure A2 reinforces our confidence that the randomization test identifies cheating.

Robustness Checks The randomization tests reported in Figure 3 assume that students only copied answers from neighbors in the same row. If students copied answers from other students sitting farther away, it goes undetected by the specification of the randomization tests. Figures A3 and A4 in the Appendix present evidence for a variety of alternative specifications (see Panels A). The figures demonstrate that in our context, copying from other students was, indeed, confined to direct neighbors within rows. Another robustness check resamples individuals within treatments (i.e., also across halls) instead of within lecture halls. Panel B in Figure A3 and A4 reports the respective results and shows that our findings are robust. We also performed the same set of robustness checks for the randomization test in the monitoring treatment. Independent of the specification, we consistently find no evidence for cheating under close monitoring.

2.2.2 Students’ Ability and Cheating

The previous subsection established that actual neighbors shared a suspiciously high number of similar answers under baseline monitoring. We expect to find especially strong traces for cheating if at least one of the two students is of low ability and was, therefore, less likely to know the solutions to the multiple-choice problems. In the following, we demonstrate, in line with studies based on student self-reports (Genereux and McLeod, 1995; McCabe and Trevino, 1997), that it is indeed the low-ability students who cheat. Particularly, the evidence for cheating is confined to pairs in which both students are of low ability, and, hence, give many wrong answers. This implies that, in our context, plagiarism mostly leads to the copying of incorrect answers and suggests that the similarity in incorrect answers is the most powerful measure of cheating.

Testing Procedure Subsequently, we suggest a simple approach to study how cheating depends on ability. As randomization tests do not allow us to

14We prefer to resample individuals within halls because this resampling scheme decreases the probability of false positives by controlling for potential hall effects. The flipside is that this scheme potentially increases the likelihood of false negatives: The counterfactual distribution could pick up other forms of plagiarism (i.e., non-row-wise plagiarism). A randomization scheme that resamples individuals within treatments takes care of this potential problem.

15In the monitoring treatment (i.e., in the absence of plagiarism), a worsening of the high-school GPA by one-standard deviation is, on average, associated with a decrease in the number of correctly solved multiple choice problems by 0.42 standard deviations.
perform such an analysis, we employ a regression-based approach.

Let us start with introducing our baseline model. The model also rests on the idea of using counterfactual neighbors as a control group for actual neighbors sitting in the same row. Defining pairs of students as the unit of observation, we consider the following simple regression without controls:

$$Y_{mp} = \beta_0 + \beta_p \cdot N_p + u_{mp}, \quad (1)$$

where $Y_{mp}$ takes a value of one if both students of a pair $p$ gave the same answer to a particular multiple-choice problem $m$. Note that $p$ can represent actual and counterfactual pairs. Further, $N_p$ indicates whether ($N_p = 1$) or not ($N_p = 0$) a pair of students consisted of actual neighbors sitting next to each other in the same row. Because we randomly assigned students to seats, $E(\beta_p)$ is a consistent reduced-form estimate of the average effect of being a pair of actual neighbors (as opposed to counterfactual ones) on the probability that both students gave the same answer. We call this the average neighbor effect (ANE). An ANE significantly larger than zero indicates cheating.

To test how cheating depends on students’ ability, we extend the model to include students’ final high-school GPA. The average high-school grade reflects a student’s performance in the final years in high school and should, therefore, be a reasonable proxy for ability. Using high-school grades, we can flexibly decompose the neighbor effect into a part that depends on the students’ abilities, and a part that does not. Formally, the decomposition reads:

$$\beta_p = E(\beta_p | B_p, W_p) + r_p$$

$$= \sum_{i=1}^{I} \beta_{2i} \cdot B_{i,p} + \sum_{j=1}^{I} \beta_{3j} \cdot W_{j,p} + \sum_{i=1}^{I} \sum_{j=1}^{I} \beta_{4ij} \cdot B_{i,p} \times W_{j,p} + r_p. \quad (2)$$

In this equation, $B_{i,p}$ reflects the high-school grade of the one student of a pair $p$ who performed better in school. Specifically, $B_{i,p}$ consists of four dummy variables indicating whether the student’s high-school grade was A, B, C, or D. Equivalently, $W_{i,p}$ are indicators for the high-school grade of the student having performed worse in high school, and $r_p$ denotes further pair-specific heterogeneity of the neighbor effect. Note that by construction, we have $E[r_p | B_p, W_p] = 0$.

To estimate the neighbor effects for different $B - W$ combinations, we simply plug (2) into (1) and add the following baseline terms to our model: $\beta_1 \cdot N_p$, $\sum_{i=1}^{I} \alpha_{2i} \cdot B_{i,p}$, $\sum_{j=1}^{I} \alpha_{3j} \cdot W_{j,p}$, and $\sum_{i=1}^{I} \sum_{j=1}^{I} \sum_{j \geq i} \alpha_{4ij} \cdot B_{i,p} \times W_{j,p}$. Our
specification then becomes:

$$Y_{mp} = \beta_0 + \sum_{i=1}^{I} \beta_{2i} \cdot B_{i,p} + \sum_{j=1}^{I} \beta_{3j} \cdot W_{j,p} + \sum_{i=1}^{I} \sum_{j \geq i}^{I} \beta_{4ij} \cdot B_{i,p} \times W_{j,p} \times N_p$$

$$+ \sum_{i=1}^{I} \alpha_{2i} \cdot B_{i,p} + \sum_{j=1}^{I} \alpha_{3j} \cdot W_{j,p} + \sum_{i=1}^{I} \sum_{j \geq i}^{I} \alpha_{4ij} \cdot B_{i,p} \times W_{j,p} + \epsilon_{mp},$$

with $\epsilon_{mp} = u_{mp} + r_p \cdot N_p$. If $N_p$ is randomized, the OLS estimators of $\beta_1$, $\beta_2$, $\beta_3$, and $\beta_4$ are unbiased and consistent. The OLS estimators of $\alpha_2$, $\alpha_3$, and $\alpha_4$ pick up potential correlations between the grade variables and $u_{mp}$. We consider observations from both exams and all halls with baseline monitoring (CONTROL and COMMITMENT). The regressions also include an exam dummy to control for potential exam effects.

Two further details of our regression-based approach are worth noting. First, as previously discussed, we expect the issue of how ability affects cheating to be related to the question of whether cheating translates into an above-normal share of identical correct or identical incorrect answers. We, therefore, estimate two different specifications: One that uses an indicator for identical correct answers as the dependent variable and one considering identical incorrect answers. Second, we construct our estimation sample such that it consists of (a) all pairs of actual direct neighbors in the same row and (b) all pairs of counterfactual neighbors (i.e., non-neighbors) who were sitting in the same row. Our regressions, hence, identify the neighbor effects by focusing on within-row variation and comparing actual neighbors in row $r$ with all the counterfactual pairs of students who were not direct neighbors but sat in the same row. This approach has two benefits. One is that it allows us to cluster standard errors at the row level. This clustering is necessary because the evidence from our randomization tests shows that direct neighbors in the same row plagiarized answers from each other. The other benefit is that because this approach indirectly controls for row effects, it is even more conservative than our previously considered randomization schemes that did not account for possible row-specific differences in cheating behavior.

**Results** Figure 4 presents the primary results of our linear probability model that either uses identical incorrect answers (Panel A) or identical correct answers (Panel B) as the outcome variable. To construct this figure, we estimate model (1) and subsequently calculate the ANEs (2) for all potential grade

---

16 For simplicity, the regression equation abstracts from the fact that due to collinearities, some of the variables are dropped when estimating the model.
combinations. The Panels A1 and B1 show the results. The horizontal axis depicts the ability (measured by the high-school grade) of the worse student and the vertical axis that of the better student. The colors in the graph indicate the size of the ANEs, ranging from the smallest neighbor effects (blue) to the largest (red). The Panels A2 and B2 show the associated p-values.\footnote{We use a thin-plate-spline interpolation to predict values for finer grade steps (e.g., for A+, B+, etc.).}

The structure of cheating emerging from Panel A is clear-cut: Cheating is confined to pairs in which both students are of low ability.\footnote{Importantly, the lack of significance for high-ability students is not reflecting low statistical power due to small sample sizes: For example, 7.5\% of all observations are pairs in which the better of the two students earned an A. We identify significant effects from comparable sample sizes at the opposite end of the grade distribution.} To clarify and elaborate on this conclusion, let us consider Figure 4 in detail. Regarding identical incorrect answers, the estimated neighbor effect (as our indicator of cheating) is the largest for pairs in which both students are of low ability. Moreover, the p-values indicate that the effect is significantly different from zero only for pairs located in the upper-right part of the colored area.

To get a sense of the effect size, consider two students whose high-school grade was C. Counterfactual pairs of this type gave, on average, identical and incorrect answers to 3.7 percent of all multiple-choice problems. Starting from this baseline probability, our model predicts a positive effect of being a pair of actual neighbors of 2.3 percentage points. In absolute terms, the average number of identical incorrect answers for symmetric counterfactual C-grade pairs was 1.1 (out of 30). The neighbor effect adds to this baseline another 0.69 identical incorrect answers (on average). If we, instead, consider symmetric pairs of students with even lower ability (high-school grade: C-), the average neighbor effect already amounts to 18 percentage points (baseline probability for counterfactual pairs: 3.8 percent). This value corresponds to an additional 5.4 identical incorrect answers, on average. We conclude that cheating among low-ability students significantly increases the likelihood of identical incorrect answers.

Panel B presents the corresponding results for jointly correct answers. The evidence is, again, unequivocal, and supports the interpretation that cheating is mainly driven by neighbor pairs in which both students are of low ability. To see this, note that copying the answers of high-ability students should increase the above-normal similarity in jointly correct answers. However, we find that the grade-specific ANEs are widely insignificant (Panel B2). Hence, we are unable to detect any significant above-normal similarity in correct answers among actual neighbors.

Overall, Figure 4 generates two insights. First, plagiarism between low-ability students, who happened to be seated next to each other, explains the above-normal similarity in neighbors’ answers under baseline monitoring.
Second, when low-ability students, who tend to give many wrong answers, copy from each other, they naturally share an above-normal number of identical incorrect answers. By contrast, there is no evidence of above-normal similarities in jointly correct answers. We conclude that identical incorrect answers are the more powerful indicator of plagiarism in our context. We, therefore, mainly focus on this measure when evaluating the effects of our treatments.

2.3 Commitment Effects

Building on the evidence presented previously, this section examines our primary topic by analyzing how the commitment request has affected cheating.

Testing Procedure The regression-based estimation strategy easily extends to the identification of treatment effects. Instead of estimating grade-combination specific neighbor effects as in equation (2), the following regressions account for treatment-specific neighbor effects. The decomposition of the neighbor effect becomes

\[ \beta_p = \beta_1 + \beta_2 \cdot C_p + \epsilon_p, \]  

where \( C_p \) denotes an indicator for the commitment treatment. As before, we plug (3) into (1), add \( \beta_3 \cdot C_p \) to the model, and obtain:

\[ Y_{mp} = \beta_0 + \left[ \beta_1 + \beta_2 \cdot C_p \right] \times N_p + \beta_3 \cdot C_p + \epsilon_{mp}, \]

with \( \epsilon_{mp} = u_{mp} + r_p \cdot N_p \). We then estimate the coefficients with OLS and cluster the standard errors at the row level. As described previously, the regressions use an indicator for identical incorrect answers as an outcome. As a robustness check, Table A1 in the Appendix presents the results for regressions that instead rely on all types of identical answers (correct and incorrect).\(^{19} \)

Results To explore the role of commitment for cheating, Table 2 reports the results of our linear probability models. We provide \( p \)-values in brackets.

\(^{19} \)Because cheating is confined to pairs in which both students tend to give wrong answers (see previous discussion), adding identical correct answers to our outcome introduces noise to the dependent variable and makes the identification of the neighbor effect more difficult. However, models that counteract this efficiency loss by adding covariates confirm our results. Note that the covariates leave the size of the main regression coefficients unchanged.
Before discussing our results, we highlight that, if two students independently and randomly picked one of the four statements, then the predicted average probability that they would have shared an incorrect answer is 0.0169. In comparison, the empirical baseline probability that two counterfactual students in the CONTROL group shared an identical incorrect answer was a bit higher than under random picking; it amounted to 0.0364. Against this backdrop, we evaluate the coefficients in Table 2.

Beginning with the unconditional estimates in Column (1), we note three points. First, the coefficient of the non-interacted treatment indicator is not statistically significant. The similarity in the answers of counterfactual pairs in the COMMITMENT treatment was not significantly different from the baseline level of 0.0364 in the CONTROL group. This result is in line with the interpretation that our experimental design successfully eliminated differences across treatments that might have affected non-cheating related correlations between students’ answers.

Second, we identify a positive neighbor effect in the CONTROL group. The coefficient for the non-interacted actual-neighbors dummy is positive and significant and amounts to 0.0073. Relative to the baseline probability of 0.0364, being a pair of actual neighbors increased the probability of an identical incorrect answer by 20%. We are, hence, able to replicate the finding of the randomization test that students cheated in the CONTROL condition.

Finally, we turn to the main result of the field experiment and evaluate the coefficient of the interaction term Commitment × Actual Neighbors: It is equal to 0.0088 and significantly different from zero. Thus, in the COMMITMENT treatment, the effect of being a pair of actual neighbors on the likelihood of providing identical incorrect answers increased relative to the CONTROL group. The increase is not only statistically significant but also substantial in size. We are unable to reject the hypothesis that the interaction effect is equal to the coefficient of Actual Neighbors (F-Test; \( p = 0.779 \)). Put differently, we cannot reject that commitment to the no-cheating rule increased the probability of an identical answer to an extent that mirrors the baseline effect of sitting next to each other.

A further question that arises from the previous analysis is whether the increase in cheating in the COMMITMENT treatment led to better grades. The evidence suggests that this was not the case. Regressing the percentage of problems solved correctly on a dummy for the COMMITMENT treatment, we consistently find, across a number of specifications, that the coefficient of

---

20The number of observations is derived as follows: Denoting with \( K \) the number of individuals in one row \( r \), we obtain \( \frac{K(K-1)}{2} \) unique pairs for this particular row (of which \( K-1 \) are unique actual neighbor pairs). The total number of observations is the sum over all pairs (considering all rows).

21This value incorporates that the average probability of a given multiple-choice question being answered correctly was 70%.

---
the treatment indicator is very small and insignificant.\textsuperscript{22} This result is in line with the previously reported evidence that pairs of low-ability students copied incorrect answers from each other, as this form of cheating does not affect the percentage of problems solved correctly.\textsuperscript{23}

**Robustness Checks** The estimations reported in Table 2, Columns (2) to (4), provide several robustness checks. Column (2) controls for multiple-choice fixed effects. Hereby, we partial out problem-specific factors that might affect the degree of similarity in neighbors’ answers (like the difficulty of the question) and identify cheating only from the within-multiple-choice problem variation. Column (3) 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 the worse student as well as interactions). Column (4) includes all the control variables. Because we randomly assigned students to seats, there is no a priori reason to expect the controls to affect the coefficients of interest. Indeed, modifying our regressions along these lines leaves the point estimates virtually unchanged.

In the Appendix, we report further versions of our estimations and present additional robustness checks. Table A2 displays the results of regressions that include an indicator variable for each hall to control for hall-specific differences in the similarity of the students’ answers. The findings are virtually unchanged. The results are also robust against estimating logit models instead of linear probability models (see Table A3 in the Appendix).\textsuperscript{24} Table A4 pools the data across all treatments (\textsc{control}, \textsc{commitment}, and \textsc{monitoring}) and reports estimations that also include a neighbor effect for the \textsc{monitoring} treatment.\textsuperscript{25} The coefficient of the interaction term \textit{Monitoring}

\textsuperscript{22}Pooling both exams and denoting the treatment indicator by $\beta_1$, we find a value of $\beta_1 = -0.181\%$ ($p$-value=0.866) in a regression without controls and $\beta_1 = -0.066\%$ ($p$-value=0.947) if we add strata variables. The $p$-values are for specifications with row clusters. To see that the effects are negligible in size, note that the average student in the \textsc{control} group answered 72.4\% of all multiple-choice questions correctly. We obtain very similar results if we additionally cluster the standard errors at the individual level or run separate regressions in both exams.

\textsuperscript{23}Our experimental design also allows us to study if students who signed the commitment request in the exam on principles of economics cheated more in the exam on principles of business administration than students who did not sign the commitment request in the first exam. We find no evidence in line with this hypothesis.

\textsuperscript{24}Prompted by the article by King and Zeng (2001), one may wonder whether our estimates are biased because of rare-event data; recall that the share of identical incorrect answers is below four percent. However, the underlying problem that causes a rare-event bias is a small number of cases on the rarer of the two outcomes. Because we have almost 5400 observations for this case, our estimations are not subject to this problem.

\textsuperscript{25}Equation (3) then becomes $\hat{\beta}_p = \beta_1 + \beta_2 \cdot C_p + \beta_3 \cdot M_p + r_p$, where $M_p$ is an indicator for
Actual Neighbors is negative and statistically significant. Furthermore, we cannot reject the hypothesis that the similarity in actual neighbors’ answers under close monitoring was equal to the similarity in the responses of counterfactual neighbors in the control group (F-Test; $p = 0.533$). This result further strengthens our confidence in the applied methods to detect cheating.

3 Suggestive Evidence on Channels

In the previous section, we have shown that the commitment treatment increased cheating. To assess the external validity of this finding, it is crucial to understand the mechanisms through which this effect operates (Deaton, 2010; Ludwig et al., 2011; Deaton and Cartwright, 2018). However, providing evidence on channels is always challenging, and in our case particularly difficult. Not only is cheating in exams a type of behavior that is, in principle, difficult to measure, but also the motives that lead individuals to cheat are, by nature, unobservable. In addition, the institutional environment of university exams limits our options to collect data that could shed light on channels. In the following, we, nevertheless, take-up the challenge of examining the forces that drive the treatment effect and present evidence on mechanisms. The evidence comes from an additional field experiment, which we implemented in the exam on principles in economics in a later year. Given the limitations we face (discussed in more detail below), we consider the evidence as being suggestive rather than being entirely conclusive.

3.1 Possible Channels

One can think of at least four channels through which commitment to a no-cheating rule could increase cheating. First, commitment requests may decrease the perceived severity of the sanctions associated with cheating. One situation in which this type of effect could occur is when students overestimate the sanction in the absence of commitment. A commitment request may then direct the students’ attention towards the real level of the sanction, decreasing its perceived severity. Second, commitment requests could also lower the perceived detection probability, for example, by signaling that monitoring of students is difficult or even impossible. Third, students could perceive the commitment request as a signal that cheating in exams is widespread, weakening the so-called perceived “descriptive norm” of academic integrity. Students with a preference for conformity with an

26 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. The considered channel is, hence, one
existing social norm (Bernheim, 1994) would then cheat more. Fourth, commitment requests may increase cheating through psychological reactance. Brehm’s (1966) theory of reactance states that individuals value their freedom of choice and become upset when facing restrictions (such as commitment requests). The triggered emotional state can induce actions to re-assert the lost freedom, for example, by engaging in the forbidden activity (boomerang effect) or another restricted activity associated with lower costs (related boomerang effect). In this vein, commitment requests that prohibit cheating could trigger not only more cheating but also a higher level of non-compliance with other exam rules.

3.2 Experimental Design

3.2.1 Perceptions

Students’ perceptions of sanctions, detection probabilities, and norms are not directly observable. The standard approach to tackle this issue is to elicit students’ perceptions using survey techniques.

Complications Two complications forbid an analysis of perceptions using our initial design. 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 an online platform used to invite students to participate in online surveys. This newly established platform allowed us to invite students who took part in the repetition of the field experiment to an online survey that measured perceptions after the exam. Second, the fact that we had to survey students after the exam complicates identification. To understand the potential issue, consider, for example, perceived norms as a channel. If the COMMITMENT treatment increases cheating, post-exam questions on norms may reflect that students in the COMMITMENT-treatment rooms observe more cheating in their lecture hall than students in the CONTROL-group rooms, instead of reflecting shifts in the perceived norm. This would lead us to overestimate the commitment effects on perceived norms. To tackle this issue, we adjusted the randomization scheme relative to the initial experiment and randomly assigned students to the COMMITMENT treatment and the CONTROL group within lecture halls when repeating the experiment. This design element ensures that students in both treatments experienced the same level of cheating by peers.

Platform and Survey We recruited students for the survey through the previously described online platform. Students registered with this platform that runs through descriptive norms.
ularly receive email invitations to participate in surveys or to perform other research-related tasks. Participants usually get a payoff that is communicated in the invitation email and paid via bank transfer.

The details of the elicitation process were as follows. Only a subsample of the students who took the exam in which we implemented the new experiment were registered with the online platform (114 out of 534 students). Two hours after the exam, we invited these individuals via email to participate in a survey after the exam. Filling out the survey questionnaire took about five minutes, and participants received a flat payoff of €3.50. 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 perceive exams at the university”. It also asked participants to think about their last exam when answering the questions. To prevent that students foresaw our goal to study the impact of the no-cheating declaration on their survey responses, we did not, however, refer to the previous exam on principles of economics at any point during the survey.

As for the survey design, Table 3 summarizes the questions on perceptions. First, we elicited the perceived sanction for cheating by requesting students to indicate what they think is 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 subjects’ beliefs about the percentage of peers who cheated in the last exam. Each question came in two versions. The first version refers 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.

One obvious limitation of our approach is the small sample size, which is due to the fact that not all students are registered users of the survey platform. Our survey sample consists of 60 students who completed the survey within two days after the invitation.27

### 3.2.2 Psychological Reactance

Because reactance is, by definition, an unpleasant motivational arousal that emerges when students commit to the no-cheating rule, it would be natural to measure its state during the exam. Given that we were unable to implement surveys within exams, we take a different approach. Our idea is to test one additional prediction of Brehm’s (1966) theory of reactance, the so-called “related boomerang effect”. Particularly, according to Brehm’s theory, the

---

27There was no other exam scheduled for freshmen students within two days after the exam in principles in economics.
students can restore their behavioral freedom not only by violating the no-cheating rule but also by actions in related (forbidden) domains. According to Brehm (1966), the related boomerang effect is particularly likely if non-compliance in the related domain comes at low or no costs.

**Related Restricted Activity** Students were required to provide their email address on the cover sheet of the exam materials (together with their student ID, name, seat number, and hall number). We test for a related boomerang effect in the repetition of the field experiment by examining the impact of the commitment request on compliance with the rule to provide an email address. Importantly, students who do not provide an email address do not incur any cost. The reason is that not providing an email address does not affect students’ grades, and email addresses are also irrelevant for the grading process. For example, neither the supervisors nor the staff members who graded the exams exam did rely on email addresses to identify students. Instead, supervisors identified students through their names and student IDs, which they also checked and verified during the exam. Furthermore, they also verified adherence to the predetermined seating plan using the seat number and the hall number.

### 3.2.3 Treatments

We evenly split the sample of students who took the exam in which we repeated our experiment into a **COMMITMENT** treatment and a **CONTROL** group, and, as previously described, randomly assigned the treatment status within exam halls. The **COMMITMENT** treatment was identical to the treatment in the initial experiment.

This type of design has two potential drawbacks. First, it forbids us to re-estimate the effect of commitment on cheating. The reason is that, in this case, neighbors end up in different treatments. Similarities in their answers, thus, reflect a mix of cheating in both treatments. Second, there might be spillovers across treatments. Such spillovers most likely equalize outcomes between treatments and, hence, tend to downward-bias the estimated effect of the commitment request on students’ survey responses and the probability of a missing email address.

We tried to reduce this downward bias by keeping the layout of the cover sheet of the exam materials identical between the **COMMITMENT** and **CONTROL** group. Particularly, instead of the no-cheating declaration, the **CONTROL** group’s cover sheet contained a text of equal length with technical information on how to handle the exam materials (see the Appendix). As a result, the exam materials looked very similar in both groups. Furthermore, in the **COMMITMENT** and the **CONTROL** group, the technical information was also printed on the second page of the exam materials. Hence, students
in both treatments received the same set of technical information. Despite these design elements, we are, nevertheless, unable to preclude any form of spillovers across treatments.

### 3.3 Commitment Effects

#### 3.3.1 Perceptions

**Sample** The survey participants have similar observable characteristics as non-participants. The only sample imbalance is that participants have slightly better high-school GPAs (see Table A6 in Appendix). Furthermore, in the sample of participants, the COMMITMENT treatment \((N = 26)\) and CONTROL group \((N = 34)\) are well-balanced in all observable characteristics (see Table A7 in Appendix).

**Results** We start our analysis by studying the perceived sanctions for copying answers or using unauthorized materials. The first result is that independent of the treatment, a vast majority of students correctly indicated that cheaters would fail the exam (copying answers: 70.0%; unauthorized materials: 86.7%). Moreover, Table 4 shows no systematic differences in the perceived sanctions between the COMMITMENT and CONTROL groups. Using Fisher’s exact tests, we cannot reject the null hypotheses that the COMMITMENT treatment did not affect the distribution of answers. Thus, in sum, we not only conclude that most of the survey participants were aware of the punishment for non-compliance but also that students who signed the commitment request did not report significantly different perceived sanctions compared to CONTROL-group individuals.

Our next step is to study the effects of the commitment request on the perceived detection probability. To that end, we use the students’ answers to questions \(D1\) and \(D2\) (for the questions, see Table 3) as outcome variables, and estimate the OLS model

\[
Y_{ih} = \gamma_0 + \gamma_C C_{ih} + X_{ih} \gamma_X + \pi_h + u_{ih},
\]

where \(Y_{ih}\) is the stated perception of student \(i\) seated in hall \(h\), \(C_{ih}\) is an indicator for the COMMITMENT treatment, and \(X_{ih}\) is a vector of student controls (age, gender, and high-school GPA). Additionally, the regression includes exam-hall fixed effects \(\pi_h\) to absorb exam-hall specific drivers of perceptions. The result tables also report heterogeneity-robust and cluster-robust \(p\)-values. To derive the latter, we use a wild cluster bootstrap procedure that accounts for the small number of clusters (Cameron et al., 2008). In addition to the treatment effects for the individual outcomes, we also report
average standardized effects according to Kling et al. (2004) and Clinging-smith et al. (2009) and exploit Mann-Whitney-U-Tests to non-parametrically test for treatment differences. Columns (1) to (3) in Table 5 show that the COMMITMENT 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 treatment effect on the perceived detection probability suggests that, at least in the sample of survey participants, the COMMITMENT treatment did not significantly shift the expected sanction for cheating.

We next analyze how the COMMITMENT treatment impacts students’ descriptive norm of academic integrity. Our analysis begins by estimating regressions in the spirit of equation (4) that use the participants’ perceived frequency of cheating as an outcome variable (see questions N1 and N2). The point estimates suggest that compared to CONTROL-group individuals, students who signed the commitment believe that about four additional peers (out of 100) plagiarized (see Column (4) in Table 5) or used unauthorized materials (see Column (5)). However, only the effect for 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 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 copying answers and using unauthorized materials.28 As expected, 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 COMMITMENT 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 also highly significant. Moreover, Column (10) displays a positive and significant average standardized effect for all four outcomes, capturing the perceived behavior of other students.

In summary, participants who signed the commitment request expected more cheating. In contrast to this result, we do not find any evidence for

28A possible concern could be that the survey respondents perceived both questions to be very similar. However, the correlation between the responses to both questions is only 0.67.
a shift in the perceived sanctions. Given the already discussed limitations that result from our inability to observe perceptions directly (small sample, measurement issues, spillovers), the survey evidence cannot ultimately identify the mediating mechanisms. The patterns in the data, however, at least suggest that the commitment request weakened the survey participants’ descriptive norms of academic integrity.

### 3.3.2 Psychological Reactance

**Sample**  In total, 534 students participated in the repetition of the field experiment.\(^{29}\) Table A8 in the Appendix shows that the COMMITMENT and the CONTROL group are balanced in observable characteristics.

**Results**  As discussed in Section 3.2, we test for the occurrence of a related boomerang effect by studying the effect of the commitment request on the probability of a missing email address. To that end, we use a version of model (4) that exploits a dummy indicating a missing email address as an outcome. We again account for hall fixed-effects and present different types of standard errors. Table 6 depicts the results. Column (1) shows that without further controls, the commitment request more than doubles the probability that a student does not provide an email address. Column (2) reports a similar effect if we include the same set of control variables that we used when estimating the treatment effects on perceptions. These findings are in line with the idea that, due to psychological reactance, the commitment request has triggered a preference for acting in opposition to the exam rules.

Further evidence helps us to benchmark the results on missing emails. Specifically, we do not find significant impacts of the commitment request on the probabilities that other student-related pieces of information on the cover sheet are missing.\(^{30}\) This finding suggests that the effect on missing emails unlikely reflects that students in the COMMITMENT group made, in general, more mistakes when completing the cover sheet.

Although the evidence on missing email addresses is in line with the concept of related-boomerang effects, we cannot know for sure whether reactance explains this effect. Because reactance is a general motivational arousal that may shape responses to commitment requests in multiple contexts, we, however, believe that further exploring the role of reactance is important. In this spirit, we initiated a companion research project that studies the impact

---

\(^{29}\)As in the original field experiment, we excluded students who had failed the exam previously and, therefore, were not taking the exam for the first time.

\(^{30}\)The detailed results are as follows: First, independent of the treatment, all students provided their Student ID, first name, and last name. Second, only one student in the CONTROL group did not fill in her seat number. Third, in the COMMITMENT treatment, 12 individuals did not provide a hall number, compared to 10 in the CONTROL group.
of commitment requests on cheating in the laboratory (Cagala et al., 2019). In line with the evidence on missing email addresses, the results of the companion paper show that individuals with particularly high values on Hong’s Psychological Reactance Scale cheat more after having signed a commitment request.

4 Conclusion

Academic cheating is a widespread and wasteful illicit activity. Therefore, educators around the world spend considerable resources to uphold academic integrity. One of the frequently used measures to fight academic cheating is to request that students commit to a no-cheating rule. However, there is no causal evidence on how such commitment requests affect academic cheating. We contribute to filling this gap in the literature by providing experimental evidence on how requesting a commitment to a no-cheating rule affects plagiarism among students. The evidence originates from a field experiment at a German university. Students registered for two exams were randomly allocated to either a CONTROL group or a COMMITMENT treatment. In the latter, they signed a declaration of compliance with the existing no-cheating rule before the beginning of the exam.

In our analysis, we focus on academic cheating taking the form of copying multiple-choice answers from neighbors. By randomly assigning students to seats, we ensure that under the null hypothesis of no cheating, the probability of choosing the same answer to a given multiple-choice question for students sitting next to each other should not be different from the probability for students who were not sitting side by side. Comparing the similarity in the answers of actual neighbors to the similarity among counterfactual neighbors allows us to identify plagiarism.

Exploiting our methods, first, we document that students plagiarize by copying answers from their neighbors, both in the COMMITMENT and the CONTROL group. Second, we show that the above-normal similarity in neighbors’ answers reflects copying among pairs of low-ability students. Third, evaluating the effect of requesting a commitment, our findings contradict the conjecture that commitment reduces cheating. To the contrary, the above-normal similarity in neighbors’ answers is significantly higher in the COMMITMENT treatment relative to the CONTROL group. This finding implies that the commitment request backfired and induced more cheating. Fourth, as for channels, we present suggestive evidence that the COMMITMENT treatment weakened the “perceived descriptive norm of academic integrity” and triggered a reactance-induced “related boomerang effect”.

We believe that our findings have implications for the fight against academic cheating. Most importantly, our main result that commitment requests
can backfire implies that educators around the world should think twice before implementing policies that require students to commit to existing no-cheating rules. Our paper may also speak to additional contexts in which commitment requests are frequently used\(^{31}\) While we cannot know whether commitment requests backfire in non-educational settings, our results certainly warrant further investigations into how effectively various forms of commitment requests induce truthful reporting in different contexts.

**References**


\(^{31}\)For example, when individuals and firms report tax-relevant information to the tax administration, they are commonly requested to sign a declaration confirming the truthfulness of the submitted information. Similarly, individuals applying for social welfare benefits and firms bidding for government contracts must submit declarations of compliance with a host of regulations. Also, all of the Fortune Global 500 corporations have a code of conduct, frequently including a declaration of compliance that newly hired staff have to sign.


Table 1: Balancing Checks

<table>
<thead>
<tr>
<th></th>
<th>Exam 1</th>
<th></th>
<th>Exam 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Commitment</td>
<td>Monitoring</td>
<td>Difference</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Gender (Female = 1)</td>
<td>0.54</td>
<td>0.56</td>
<td>0.50</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.05)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>High-School GPA</td>
<td>2.47</td>
<td>2.48</td>
<td>2.50</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Math Proficiency</td>
<td>0.75</td>
<td>0.73</td>
<td>0.73</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Field of Study (Econ. &amp; Sociology = 1)</td>
<td>0.07</td>
<td>0.06</td>
<td>0.09</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Age</td>
<td>19.6</td>
<td>19.6</td>
<td>19.6</td>
<td>-0.04</td>
</tr>
<tr>
<td>Bavaria</td>
<td>0.81</td>
<td>0.83</td>
<td>0.84</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>333</td>
<td>208</td>
<td>225</td>
<td>204</td>
</tr>
</tbody>
</table>

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 Commitment 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.
Table 2: Responses to Commitment: Actual and Counterfactual Pairs

<table>
<thead>
<tr>
<th></th>
<th>(1) Unconditional Estimates</th>
<th>(2) Multiple-Choice Controls</th>
<th>(3) Pair Controls</th>
<th>(4) All Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commitment</td>
<td>0.0008</td>
<td>0.0008</td>
<td>0.0002</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>[0.7997]</td>
<td>[0.7999]</td>
<td>[0.9316]</td>
<td>[0.9317]</td>
</tr>
<tr>
<td>Actual Neighbors</td>
<td>0.0073***</td>
<td>0.0073***</td>
<td>0.0066***</td>
<td>0.0066***</td>
</tr>
<tr>
<td></td>
<td>[0.0012]</td>
<td>[0.0012]</td>
<td>[0.0016]</td>
<td>[0.0016]</td>
</tr>
<tr>
<td>Commitment × Actual Neighbors</td>
<td>0.0088**</td>
<td>0.0088**</td>
<td>0.0081**</td>
<td>0.0081**</td>
</tr>
<tr>
<td></td>
<td>[0.0190]</td>
<td>[0.0189]</td>
<td>[0.0126]</td>
<td>[0.0125]</td>
</tr>
<tr>
<td>Pair Controls</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Share of Identical Incorrect Answers amongst Counterfactual Pairs in Control Group: 0.0364

Number of Clusters: 81
Number of Observations: 140,937

Notes: This table reports estimates of the treatment-specific effect of being a pair of actual neighbors on the probability that two paired students provide identical incorrect answers. Estimates are based on linear probability models. Column (1) presents the unconditional estimates. Column (2) controls for multiple-choice fixed effects. Column (3) 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 (4) includes all the aforementioned control variables. All specifications include an exam dummy. Standard errors are clustered at the row level; p-values in brackets.
Table 3: 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.

Table 4: Post-Exam Survey: Perceived Sanctions

<table>
<thead>
<tr>
<th>Perceived Sanction: Copying</th>
<th>Control</th>
<th>Commitment</th>
<th>Perceived Sanction: Unauthorized Materials</th>
<th>Control</th>
<th>Commitment</th>
</tr>
</thead>
<tbody>
<tr>
<td>No sanction at all</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Verbal warning</td>
<td>8.8</td>
<td>15.4</td>
<td>5.9</td>
<td>3.9</td>
<td>3.9</td>
</tr>
<tr>
<td>Exam committee hears case and decides</td>
<td>17.6</td>
<td>15.4</td>
<td>8.8</td>
<td>7.7</td>
<td></td>
</tr>
<tr>
<td>Student fails exam</td>
<td>73.5</td>
<td>65.4</td>
<td>85.3</td>
<td>88.5</td>
<td></td>
</tr>
<tr>
<td>Student is expelled from university</td>
<td>0.0</td>
<td>3.9</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>

*p*-value, Fisher’s exact test: [0.591] [1.000]

Number of Observations: 34 26 34 26

**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 COMMITMENT treatment affected the distributions of answers. We report the corresponding *p*-values [in brackets].
Table 5: Post-Exam Survey: Treatment Effects

<table>
<thead>
<tr>
<th>Detection Probability Copying Effect (1)&amp;(2)</th>
<th>Detection Probability Unauthorized Materials Effect (3)</th>
<th>Average Stand. Students Copying Effect (4)</th>
<th>% Other Students Using Unauthorized Materials Effect (5)</th>
<th>% Others Using Unauthorized Materials Effect (6)</th>
<th>Average Social Norm Copying Effect (7)</th>
<th>Social Norm Unauthorized Materials Effect (8)</th>
<th>Average Social Norm Effect (7)&amp;(8)</th>
<th>Average Social Norm Effect (4),(5),(7),(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect of Commitment</td>
<td>-0.2</td>
<td>1.5</td>
<td>0.02</td>
<td>4.0</td>
<td>4.2</td>
<td>0.63</td>
<td>20.9</td>
<td>19.1</td>
</tr>
<tr>
<td>p-value, robust</td>
<td>[0.983]</td>
<td>[0.878]</td>
<td>[0.936]</td>
<td>[0.209]</td>
<td>[0.036]**</td>
<td>[0.040]**</td>
<td>[0.009]***</td>
<td>[0.002]**</td>
</tr>
<tr>
<td>p-value, hall cluster, wild bootstrap</td>
<td>[0.918]</td>
<td>[0.859]</td>
<td>[0.278]</td>
<td>[0.022]**</td>
<td>[0.016]**</td>
<td>[0.093]*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p-value, Mann-Whitney-U-Test</td>
<td>[0.637]</td>
<td>[0.559]</td>
<td>[0.150]</td>
<td>[0.024]**</td>
<td>[0.009]***</td>
<td>[0.042]**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control Group Mean</td>
<td>27.5</td>
<td>28.4</td>
<td>6.6</td>
<td>3.9</td>
<td>47.2</td>
<td>45.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: The table reports the effects of the COMMITMENT 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-t 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 variable 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 Appendix for the exact wording of the survey questions.
**Table 6: Testing for Psychological Reactance: Missing Email Addresses**

<table>
<thead>
<tr>
<th>Effect of Commitment</th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value, robust</td>
<td>[0.009]</td>
<td>[0.015]</td>
</tr>
<tr>
<td>p-value, hall cluster, wild bootstrap</td>
<td>[0.007]</td>
<td>[0.007]</td>
</tr>
<tr>
<td>Control Group Mean</td>
<td>0.052</td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>534</td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** The table reports the effects of the commitment treatment on the degree of psychological reactance. As a proxy for psychological reactance, we use an indicator for a missing email address on the cover sheet of the exam materials. The estimates are derived from OLS regressions. Column (1) accounts for exam-hall fixed effects. Column (2) in addition controls for gender, age, and high-school GPA. 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-t procedure that accounts for the small number of cluster (Cameron et al., 2008).

**Figure 1: Overview of Field-Experimental Design**

**Notes:** This figure visualizes the experimental design. The field experiment was implemented in two written exams. Exam 1 comprised a control group and two treatment groups, commitment and monitoring. Students assigned to the control group in Exam 1 were also sampled for the intervention in Exam 2, comprising a control group and a commitment treatment group. The figure indicates, for each treatment, the number of students assigned to the respective treatment group, and the number of students who actually took the exam. Differences between the two figures are due to the fact that students could postpone participation to later semesters.
Figure 2: Responses to a Selected Multiple-Choice Problem in One Lecture Hall

Notes: This figure shows the students’ answers to one multiple-choice problem in one specific 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.

Figure 3: Randomization Tests: Cheating by Treatment Group

A: Control

B: Commitment

Notes: This figure shows the results for the treatment-specific randomization tests. In each panel, the vertical line represents the test statistic derived from the actual seating arrangement. The bell-shaped curves show the mean-centered distributions of the test statistic under the null of no cheating on the basis of Epanechnikov kernels. For both panels, we obtain $p \leq 0.001$ (two-tailed tests).
Figure 4: Cheating: Heterogeneity with Respect to Students’ Ability

Dependent Variable A: Indicator for Identical Incorrect Answers

A1: Neighbor Effects

A2: p-values

Dependent Variable B: Indicator for Identical Correct Answers

B1: Neighbor Effects

B2: p-values

Notes: This figure examines how students’ ability (proxied by high-school performance) relates to their cheating behavior. To construct this figure, we exploit the linear probability model (1) to estimate the effect of being a pair of actual neighbors on the probability that two paired students give identical incorrect answers (Panel A) or identical correct answers (Panel B). We allow for average-neighbor-effect heterogeneity in students’ ability (see equation 2). Panels A1 and B1 demonstrate this heterogeneity: The horizontal (vertical) axis shows the grade of the worse (better) student of a particular pair, the colors indicating the size of the average neighbor effect for a specific grade combination. High-school grade combinations with the largest (the smallest) neighbor effects are colored in red (blue). The grade scale ranges from A (best grade) to D (worst grade). Panels A2 and B2 show the associated p-values. We use a thin-plate-spline interpolation to predict values for intermediate grades and cluster standard errors at the row level.
Table A1: Responses to Commitment: All Identical Answers

<table>
<thead>
<tr>
<th></th>
<th>Dependent Variable: Indicator for Identical Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) Unconditional Estimates (2) Multiple-Choice Controls (3) Pair Controls (4) All Controls</td>
</tr>
<tr>
<td>Commitment</td>
<td>−0.0029 [0.8323]                                  −0.0029 [0.8335] −0.0001 [0.9896] −0.0001 [0.9908]</td>
</tr>
<tr>
<td>Actual Neighbors</td>
<td>0.0144*** [0.0028]                                0.0144*** [0.0028] 0.0120** [0.0205] 0.0120** [0.0206]</td>
</tr>
<tr>
<td>Commitment × Actual Neighbors</td>
<td>0.0105 [0.1561]                               0.0105 [0.1567] 0.0143* [0.0684] 0.0143* [0.0683]</td>
</tr>
<tr>
<td>Multiple-Choice FE</td>
<td>No                                                 Yes                                             No                                              Yes</td>
</tr>
<tr>
<td>Pair Controls</td>
<td>No                                                 No                                              Yes                                             Yes</td>
</tr>
<tr>
<td>Share of Identical</td>
<td></td>
</tr>
<tr>
<td>Answers amongst Counterfactual Pairs in Control Group</td>
<td></td>
</tr>
<tr>
<td>Number of Clusters</td>
<td>81</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>140,937</td>
</tr>
</tbody>
</table>

Notes: This table reports estimates of the treatment-specific effect of being a pair of actual neighbors on the probability that two paired students provide identical answers (correct or incorrect). Estimates are based on linear probability models. Column (1) presents the unconditional estimates. Column (2) controls for multiple-choice fixed effects. Column (3) 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 (4) includes all the aforementioned control variables. All specifications include an exam dummy. Standard errors are clustered at the row level; p-values in brackets.
Table A2: Responses to Commitment: Results for Specifications with Room Effects

<table>
<thead>
<tr>
<th></th>
<th>(1) Unconditional Estimates</th>
<th>(2) Multiple-Choice Controls</th>
<th>(3) Pair Controls</th>
<th>(4) All Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Neighbors</td>
<td>0.0077***</td>
<td>0.0077***</td>
<td>0.0069***</td>
<td>0.0069***</td>
</tr>
<tr>
<td></td>
<td>[0.0003]</td>
<td>[0.0003]</td>
<td>[0.0009]</td>
<td>[0.0009]</td>
</tr>
<tr>
<td>Commitment × Actual Neighbors</td>
<td>0.0086**</td>
<td>0.0086**</td>
<td>0.0082**</td>
<td>0.0082**</td>
</tr>
<tr>
<td></td>
<td>[0.0213]</td>
<td>[0.0212]</td>
<td>[0.0107]</td>
<td>[0.0106]</td>
</tr>
<tr>
<td>Room FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Multiple-Choice FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pair Controls</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Share of Identical Incorrect Answers amongst Counterfactual Pairs in Control Group</td>
<td>0.0364</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Clusters</td>
<td>81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>140,937</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: This table reports estimates of the treatment-specific effect of being a pair of actual neighbors on the probability that two paired students provide identical incorrect answers. Estimates are based on linear probability models. Column (1) includes hall indicators to control for hall effects. Column (2) additionally controls for multiple-choice fixed effects. Column (3) 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 (4) includes all the aforementioned control variables. All specifications include an exam dummy. Standard errors are clustered at the row level; p-values in brackets.
Table A3: Responses to Commitment: Results of Logit Models

<table>
<thead>
<tr>
<th></th>
<th>(1) Unconditional Estimates</th>
<th>(2) Multiple-Choice Controls</th>
<th>(3) Pair Controls</th>
<th>(4) All Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commitment</td>
<td>0.0233 [0.7987]</td>
<td>0.0241 [0.8001]</td>
<td>0.0073 [0.9232]</td>
<td>0.0081 [0.9182]</td>
</tr>
<tr>
<td>Actual Neighbors</td>
<td>0.1896*** [0.0005]</td>
<td>0.1977*** [0.0005]</td>
<td>0.1697*** [0.0007]</td>
<td>0.1792*** [0.0007]</td>
</tr>
<tr>
<td>Commitment × Actual Neighbors</td>
<td>0.1848** [0.0226]</td>
<td>0.1953** [0.0224]</td>
<td>0.1705** [0.0213]</td>
<td>0.1823** [0.0206]</td>
</tr>
<tr>
<td>Multiple-Choice FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pair Controls</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Number of Clusters: 81
Number of Observations: 140,937

Notes: This table reports estimates of the treatment-specific effect of being a pair of actual neighbors on the probability that two paired students provide identical incorrect answers. Estimates are based on logit models. Column (1) presents the unconditional estimates. Column (2) controls for multiple-choice fixed effects. Column (3) 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 (4) includes all the aforementioned control variables. All specifications include an exam dummy. Standard errors are clustered at the row level; p-values in brackets.
Table A4: Commitment and Monitoring: Actual and Counterfactual Pairs

<table>
<thead>
<tr>
<th></th>
<th>Dependent Variable: Indicator for Identical Incorrect Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) Unconditional Estimates</td>
</tr>
<tr>
<td>Monitoring</td>
<td>0.0033</td>
</tr>
<tr>
<td></td>
<td>[0.4445]</td>
</tr>
<tr>
<td>Commitment</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td>[0.7989]</td>
</tr>
<tr>
<td>Actual Neighbors</td>
<td>0.0073**</td>
</tr>
<tr>
<td></td>
<td>[0.0009]</td>
</tr>
<tr>
<td>Monitoring $\times$ Actual Neighbors</td>
<td>$-$0.0095**</td>
</tr>
<tr>
<td></td>
<td>[0.0240]</td>
</tr>
<tr>
<td>Commitment $\times$ Actual Neighbors</td>
<td>0.0088**</td>
</tr>
<tr>
<td></td>
<td>[0.0176]</td>
</tr>
</tbody>
</table>

Multiple-Choice FE | No | Yes | No | Yes  
Pair Controls | No | No | Yes | Yes  

Share of Identical Incorrect Answers amongst Counterfactual Pairs in Control Group | 0.0364  
Number of Clusters | 148  
Number of Observations | 149,776  

Notes: This table reports estimates of the treatment-specific effect of being a pair of actual neighbors on the probability that two paired students provide identical incorrect answers. Estimates are based on linear probability models. Column (1) presents the unconditional estimates. Column (2) controls for multiple-choice fixed effects. Column (3) 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 (4) includes all the aforementioned control variables. All specifications include an exam dummy. Standard errors clustered at the row level, $p$-values in brackets.
Table A5: Monitoring Intensity by Lecture Hall

<table>
<thead>
<tr>
<th>Hall</th>
<th>Control Students per Supervisor</th>
<th>Commitment Students per Supervisor</th>
<th>Monitoring Students per Supervisor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>51.3</td>
<td>56.5</td>
<td>9.2</td>
</tr>
<tr>
<td>2</td>
<td>49.8</td>
<td>47.5</td>
<td>8.5</td>
</tr>
<tr>
<td>3</td>
<td>38.0</td>
<td>44.5</td>
<td>8.0</td>
</tr>
<tr>
<td>4</td>
<td>29.0</td>
<td>30.0</td>
<td>8.0</td>
</tr>
</tbody>
</table>

Treatment-specific Averages

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Commitment</th>
<th>Monitoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>46.4</td>
<td>46.6</td>
<td>8.4</td>
</tr>
</tbody>
</table>

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, the COMMITMENT, and the MONITORING TREATMENT, respectively (weights: number of students in lecture hall).
Table A6: Post-Exam Survey: Characteristics of Participants and Non-Participants

<table>
<thead>
<tr>
<th></th>
<th>Non-Participants (1)</th>
<th>Participants (2)</th>
<th>Difference (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (Female = 1)</td>
<td>0.51</td>
<td>0.52</td>
<td>-0.01 (0.07)</td>
</tr>
<tr>
<td>High-School GPA</td>
<td>2.50</td>
<td>2.34</td>
<td>0.16 (0.08)</td>
</tr>
<tr>
<td>Math Proficiency</td>
<td>2.68</td>
<td>2.53</td>
<td>0.16 (0.15)</td>
</tr>
<tr>
<td>Field of Study (Econ. &amp; Sociology = 1)</td>
<td>0.12</td>
<td>0.10</td>
<td>0.02 (0.04)</td>
</tr>
<tr>
<td>Age</td>
<td>21.3</td>
<td>21.0</td>
<td>0.39 (0.43)</td>
</tr>
<tr>
<td>Bavaria</td>
<td>0.91</td>
<td>0.88</td>
<td>0.02 (0.04)</td>
</tr>
</tbody>
</table>

| Number of Observations | 474 | 60 |

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 351 out of the 474 Non-participants and 50 out of 60 Participants.
Table A7: Post-Exam Survey: Balancing Checks

<table>
<thead>
<tr>
<th></th>
<th>Control (1)</th>
<th>Commitment (2)</th>
<th>Difference (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (Female = 1)</td>
<td>0.47</td>
<td>0.58</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.13)</td>
</tr>
<tr>
<td>High-School GPA</td>
<td>2.41</td>
<td>2.25</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.16)</td>
</tr>
<tr>
<td>Math Proficiency</td>
<td>2.63</td>
<td>2.40</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.31)</td>
</tr>
<tr>
<td>Field of Study (Econ. &amp; Sociology = 1)</td>
<td>0.06</td>
<td>0.15</td>
<td>-0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.08)</td>
</tr>
<tr>
<td>Age</td>
<td>20.9</td>
<td>21.0</td>
<td>-0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.50)</td>
</tr>
<tr>
<td>Bavaria</td>
<td>0.82</td>
<td>0.96</td>
<td>-0.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.08)</td>
</tr>
</tbody>
</table>

Number of Observations 34 26

Notes: This table shows balancing checks for the sample of post-exam online survey participants. Column (3) shows the difference in means between Commitment and Control with standard errors in parentheses. Math proficiency is only available for 50 students (27 in Control and 23 in Commitment).
<table>
<thead>
<tr>
<th></th>
<th>Control (1)</th>
<th>Commitment (2)</th>
<th>Difference (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (Female = 1)</td>
<td>0.49</td>
<td>0.53</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.04)</td>
</tr>
<tr>
<td>High-School GPA</td>
<td>2.52</td>
<td>2.45</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.05)</td>
</tr>
<tr>
<td>Math Proficiency</td>
<td>2.80</td>
<td>2.52</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.10)</td>
</tr>
<tr>
<td>Field of Study (Econ. &amp; Sociology = 1)</td>
<td>0.12</td>
<td>0.11</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.03)</td>
</tr>
<tr>
<td>Age</td>
<td>21.2</td>
<td>21.4</td>
<td>-0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.27)</td>
</tr>
<tr>
<td>Bavaria</td>
<td>0.90</td>
<td>0.90</td>
<td>-0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.26)</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>268</td>
<td>266</td>
<td></td>
</tr>
</tbody>
</table>

Notes: This table shows balancing checks for the sample of post-exam online survey participants. Column (3) shows the difference in means between commitment and control with standard errors in parentheses.
Notes: This figure is a stylized illustration of baseline monitoring (CONTROL group and COMMITMENT treatment) and close monitoring (MONITORING treatment). Gray dots represent students; black squares represent supervisors. The average monitoring intensities were 44.2 students per supervisor under baseline monitoring, and 8.4 students per supervisor under close monitoring.
Figure A2: Randomization Test: Cheating in the Monitoring Treatment

Notes: This figure shows the result of the randomization test in the monitoring treatment. The vertical line represents the test statistic derived from the actual seating arrangement. The bell-shaped curve shows the mean-centered distributions of the test statistic under the null of no cheating on the basis of Epanechnikov kernels. We obtain $p = 0.999$ (two-tailed test).
Figure A3: 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)

Control
Commitment

Norm. Share of Identical Answers

Notes: This figure examines the spatial structure of cheating (Panel A) and tests the robustness of our results with respect to the randomization schemes (Panel B). The figure also shows a sketch of a representative seating plan, in which 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 Panels A3 and B3 test for column-wise cheating (copying from 3 and 13). The Panels 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% confidence bands for the counterfactual distributions (blue spikes), and the empirical value of the relevant test statistic (red circles).
Figure A4: 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)

Notes: This figure examines the spatial structure of cheating (Panel A) and tests the robustness of our results with respect to the randomization schemes (Panel B). The figure also shows a sketch of a representative seating plan, in which the yellow circle represents a particular student who can copy answers from her neighbors 1 to 15. The Panels A1 and B1 assume that the student only copied answers from the student in seat 3. The Panels 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% confidence bands for the counterfactual distributions (blue spikes), and the empirical value of the relevant test statistic (red circles).
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
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
First Name    Seat Number
Matriculation Number    Room
Email Address

Please carefully read the information provided on the back page!