Commitment to a No-Cheating Rule Can Increase Cheating

Tobias Cagala, Ulrich Glogowsky, Johannes Rincke*

September 3, 2018

Abstract
Public institutions often require firms and individuals to commit to truthful reporting by signing a no-cheating declaration. We test different commitment requests in the laboratory and find diverging effects. Signing a no-cheating declaration decreases misreporting if it is morally charged, does not affect behavior if it is morally neutral, and backfires if it is neutral and threatens to punish. Heterogeneity analyses suggest that the backfiring effect is due to psychological reactance. Furthermore, the adverse effect is not a laboratory artifact: In a field experiment on cheating in exams, we demonstrate that students who signed a no-cheating declaration plagiarized more.

Keywords: commitment; compliance; cheating; lying; reactance

*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 Imran Rasul, Joel Slemrod, and seminar participants at various places. We are grateful for financial support from the Emerging Field Initiative of the University of Erlangen-Nuremberg. 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

For many public institutions, ranging from public schools and universities to the administration of tax, transfer, and judiciary systems, it is crucial that agents report private information truthfully. A large body of literature since Becker (1968) has shown that, in principle, deterrence can induce honest reporting. For instance, third-party information reporting reduces tax evasion of individuals and firms (Kleven et al., 2011; Kleven, 2014; Pomeranz, 2015) and close monitoring curbs the mismanagement of publicly funded institutions (Reinikka and Svensson, 2005; Olken, 2007; Ferraz and Finan, 2008, 2011; Bjorkman and Svensson, 2010). However, public institutions that want to introduce third-party reporting or close monitoring often face practical or legal constraints or may find that the costs associated with these policies outweigh the benefits. As a result, they are frequently left with non-deterrent policies to induce truthful reporting. A widely used instrument in such contexts is requesting the agent to sign a declaration of compliance with a specific set of rules. For example, many universities require students to sign an honor code or a no-cheating declaration in which students commit to academic integrity. Similarly, when individuals and firms report tax-relevant information, the tax administration commonly requests them to sign a declaration confirming the truthfulness of the submitted information.\footnote{One exemplary country that made use of such a commitment request is Sweden. Before 2002, the Swedish income-tax-return form included the following statement that individuals had to sign: “I promise in honor that the submitted figures are correct and truthful.”} Also, individuals who apply for social welfare benefits and firms bidding for governments contracts must submit declarations of compliance with a host of regulations.\footnote{Commitment requests are, of course, also used by private institutions. For example, all of the Fortune Global 500 corporations have a code of conduct, which frequently includes a declaration of compliance that newly hired staff have to sign. The compliance database of the University Houston collects the code of conducts for these corporations (see Link).}

One can think of various channels through which commitment to a no-cheating rule could affect truthful behavior positively. For example, as documented by Gneezy (2005), Erat and Gneezy (2012), and others, many agents’ reporting behavior is not only shaped by extrinsic motivations but also by an intrinsic disutility of cheating. In this spirit, commitment may increase the perceived disutility of cheating, shifting the tradeoff between truthful reporting and cheating towards more honesty. One potential reason for such a shift is that committing to a no-cheating rule can raise subjects’ attention to their moral standards (Mazar et al., 2008). Likewise, commitments could reduce cheating due to a disutility of breaking a promise (Ellingsen and Johannesson, 2004; Charness and Dufwenberg, 2006), or it could remind subjects
of an existing social norm not to cheat and thereby trigger more compliant behavior.

Working through a more indirect channel, the act of commitment could also lead agents to update their beliefs about the detection probability and the sanction associated with breaking the rule. Depending on whether the direction of the update is positive or negative, this could induce more or less truthful reporting.\footnote{Given that there is no role for commitment in a situation with perfect enforcement, it is not unlikely that individuals perceive a commitment request as a signal of imperfect enforcement. Dwenger et al. (2016) suggest a related explanation for an adverse effect of compliance rewards in a zero-deterrence setting.}

A further potentially important channel for a negative impact of commitment on truthful reporting could result from psychological reactance. Going back to Brehm (1966), the theory of reactance states that individuals have a fundamental need for behavioral freedom. This need is activated whenever individuals feel a restriction put on their options or actions, leading them to an emotional state characterized by the wish to regain their freedoms through engaging in the restricted activity. In this vein, commitment requests that imposed behavioral restrictions could lead individuals to choose deliberately the type of behavior that the request marks as (socially) undesirable.

With competing conceptual frameworks predicting very different effects of commitment, it is unclear a priori whether requesting commitment induces more or even less truthful reporting. Against this backdrop, this paper presents experimental evidence from the laboratory and the field to demonstrate how different forms of commitment affect cheating behavior. While commitments may shape behavior through all the previously described channels, in this paper, we focus on the intrinsic-motivation channels as the channels of primary interest (i.e., psychological costs of cheating and psychological reactance). Furthermore, we analyze the effects of commitment in two different types of settings.\footnote{The field experiment (Section 4) was conducted before the laboratory experiment (Section 3). Because the evidence from the laboratory allows not only for the estimation of treatment effects but also for an analysis of the channels through which commitment requests affect cheating, we nevertheless present the evidence from the laboratory first. This, in turn, creates a closer link to our conceptual framework, also presented in Section 3.}

First, we abstract from specific public institutions and use a simple cheating game inspired by Abeler et al. (2018) to demonstrate that under the controlled conditions of the laboratory, different types of commitment requests can have vastly different effects. While a commitment to a morally charged norm reduced cheating in the experiment, a morally neutral commitment request that made the loss of options under commitment salient and threatened to punish triggered psychological reactance and increased cheating.

Second, we consider universities as an important class of public institutions that frequently employ commitment requests. In a natural field experiment on cheating
in university exams, we demonstrate that the adverse effect of commitment is not an artifact of the laboratory setting. In the experiment, we randomly allocated students to a control group and a commitment treatment. All students were subject to the same monitoring conditions and the same no-cheating rule that the supervisors publicly announced before the exam. The control group featured standard exam conditions at the university without any form of commitment to the no-cheating rule. In the commitment treatment, we requested that students sign a morally neutral declaration of compliance with the no-cheating rule. Students then worked on the exam. To measure the extent of cheating in the form of plagiarism among neighbors, we exploit spatial randomization schemes to identify above-normal similarity in the answers of students sitting next to each other. The evidence reveals that, on average, students in the commitment condition cheated significantly more relative to the control condition. Hence, also in the field, a commitment request that highlights restrictions on individual behavior backfires and triggers more cheating.\footnote{We also implemented a third treatment that imposed close monitoring of students during the exam (but no commitment). The purpose of having the monitoring treatment was 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 following, we mainly report the results for the control group and the commitment treatment. We refer the reader to the Appendix for a full collection of results including the monitoring treatment.}

Our paper adds to an extensive literature on lying and cheating behavior (see Kajackaite and Gneezy, 2017 and Abeler et al., 2018 for recent contributions and additional references). In particular, we contribute to the experimental literature that studies how commitment to truthful reporting affects cheating. As discussed in the following section, this literature has mostly analyzed morally charged forms of commitment referring to an honor code (Mazar et al., 2008) or a truth-telling oath (Jacquemet et al., 2018). The main takeaway from the existing literature is that these types of commitment increase honesty. We advance this literature in three directions.

First, we use three different commitment requests to test how commitment affects truthful reporting. The treatments aim to either make ethics more salient or to shift the degree of reactance that is triggered. Our first treatment features a morally charged commitment designed to work as a moral reminder. Specifically, the treatment requests subjects to sign a no-cheating declaration referring to the “principles of ethically sound behavior”. The second treatment employs a morally neutral commitment that does not refer to any ethically loaded norm but formulates a specific behavioral restriction. We expect this treatment to be a weaker moral reminder. The goal of the third treatment is to trigger reactance: It combines the neutral com-
mitment with a threat that non-compliance will be sanctioned and hence conveys a very direct message that the behavioral freedom of subjects is restricted. We designed this declaration with reference to the literature arguing that reactance tends to increase in the severity of the threat (Pennebaker and Sanders, 1976; Brehm and Brehm, 1981; Miron and Brehm, 2006; Steindl et al., 2015). We test our treatments against a control group with no commitment and find that only the morally charged commitment request reduced cheating. For the morally-neutral commitment request (second treatment), we find no significant difference relative to the control group. In contrast, the treatment adding a sharp reactance trigger to the neutral commitment request significantly increased cheating relative to the control group.

The second novelty of our study is that we explicitly consider reactance as a channel through which commitment can affect cheating. For that purpose, we use survey data collected before the experiment to elicit whether subjects are of a more or less reactant type. We then contrast the treatment response of subjects who are particularly prone to “pushing back” if their freedom of choice is restricted to the response of subjects less prone to do so. We find that only the reactance-prone subjects in the sample cheat more if the commitment request contains a trigger of reactance. This suggests that reactance is indeed the driving force behind the backfiring effect of commitment in the respective treatment group.

A third addition to the literature lies in the fact that we provide evidence on the external validity of our main finding. The context we study with our natural field experiment is economically significant, as cheating in exams could distort the education-related signals of prospective workers and managers about their productivity. This distortion could critically impact the hiring decisions of employers.

While testing the external validity of our results is important, switching to a field context complicates the identification of cheating. We focus on cheating in the form of copying multiple choice answers from neighbors and tackle the identification problem by seating students according to a randomized seating plan. This ensures that under the null hypothesis of no cheating, the probability of choosing the same answer to a given multiple choice question for students who were sitting next to each other (and could, therefore, copy from each other) should not be different than for students who were not sitting next to each other (and could, therefore, not copy from each other). We propose two tests for plagiarism in exams that exploit this feature: a non-parametric randomization test for above-normal similarity in neighbors’ answers and a regression-based test that allows us to evaluate treatment

Jacquemet et al. (2017) mention reactance as a possible response to commitment requests, but we are not aware of any prior attempt to analyze if this channel is empirically relevant.
effects. Both tests identify cheating by comparing 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). The regression-based approach reveals that students who were requested to sign the no-cheating rule cheated significantly more: The neighbor effect, defined as the change in the probability of an identical incorrect answer among actual neighbors relative to non-neighbors, is positive and twice as large in the commitment treatment as compared to the control group.

The remainder of the paper is organized as follows. Section 2 discusses the relevant literature on how commitment affects cheating behavior. In Section 3, we present a conceptual framework, explain the laboratory experiment and discuss the findings from the laboratory. Section 4 describes the design, our identification strategy and the results of the field experiment. Section 5 concludes.

2 Literature Review

In the following, we discuss the related literature on how commitment affects unethical behavior. An extensive literature, mainly in the field of psychology, has discussed the effects of commitment. Most of the research follows the idea that commitment acts as a self-identity prime (Kettle and Häubl, 2011). In theory, evoking the self can induce honesty because it leads the subject to compare personal conduct to the socially desirable conduct addressed by the cue (Duval and Wicklund, 1972). To the extent that people strive to see themselves as ethical (e.g., Steele, 1988; Monin and Jordan, 2009; Bryan et al., 2013), the comparison induces compliance (Chou, 2015).

In fact, many studies suggest a negative association between cues that evoke the self and socially undesirable behavior like dishonesty. Some of the cues impose subtle manipulations like, for instance, different phrasings in the description of the study context. The work of Bryan et al. (2013) is one example, showing that instructions that frame cheating behavior as an undesirable identity (“being a cheater” versus “cheating”) lower the probability of cheating. Other articles that implemented similar subtle changes demonstrate that subjects cheat less after being primed by the task to write down the ten commandments (Mazar et al., 2008) or to read an honor code (Shu et al., 2011).\footnote{Invoking the self appears also to affect behavior in other, non-cheating related contexts, including voting behavior (Bryan et al., 2011) and behavior in dictator games (Haley and Fessler, 2005; Rigdon et al., 2009).}
More closely related to our paper are, however, previous studies that aim at evoking the self by requesting individuals to sign an honor code or a code of conduct. Conceptually, because individuals strongly associate their signature with their identity, signing one’s name can be thought of as a particularly strong self-identity prime (Kettle and Häubl, 2011). Indeed, such commitments seem to be powerful. For example, a prominent contribution of Mazar et al. (2008) demonstrates that, despite a punishment probability of zero, students who signed the declaration “I understand that this short survey falls under the [university] honor system” did not make use of a cheating opportunity to increase their payoff. Shu et al. (2012) report related evidence, pointing out that signing at the beginning rather than at the end of a self-report reduces cheating. Furthermore, the paper of Chou (2015) suggests that a handwritten signature reduces cheating relative to an electronic signature.\footnote{Others have also studied the effects of signing a declaration of commitment in non-cheating related contexts. See, for instance, Dickerson et al. (1992) on water conservation behavior.}

In addition to the above-mentioned literature on commitment requests, others have studied how the voluntary decision to take an oath affects dishonesty. For example, Jacquemet et al. (2018) demonstrate that subjects who signed an oath that (a) includes an explicit moral reminder and (b) frames false reporting as a lie are more likely to tell the truth. Similarly, taking an oath also increases compliance in a tax evasion experiment (Jacquemet et al., 2017).

Collecting the evidence on the effects of commitment, we note that all the discussed articles highlight the positive side of such identity primes. In contrast, the central theme of our paper is the potential drawbacks of commitment, a topic that the literature has widely neglected so far.\footnote{One exception is a randomized controlled trial from the U.K. in which researchers manipulated forms to apply for a tax discount (Behavioural Insights Team, 2012). In particular, they moved an honesty declaration from the bottom (control group) to the top of the form (treatment group). The authors expected to observe a lower application rate in the control group, which they would have interpreted as evidence of reduced fraud. Instead, the intervention increased the number of requests for the discount.} In particular, we primarily focus on psychological reactance as a specific channel through which commitment requests can increase cheating. The theory of psychological reactance assumes that subjects value certain freedoms concerning their behavior (Brehm, 1966). Whenever a particular choice is restricted, or a policy (such as a commitment request) aims at restricting it, the individual will be “directed toward the re-establishment of whatever freedom has been lost or threatened” (Brehm, 1966). Reactance, hence, operates by making the restricted behavior or object more desirable (Brehm, 2009). In economic terms, reactance provides individuals with additional utility from pursuing the restricted action.
Empirically, the psychological literature has shown that imposing restrictions on behavior can indeed affect an individual’s attitudes towards the restricted activity and might lead to increased engagement in this behavior (see, e.g., the reviews of Miron and Brehm, 2006; Rains, 2013; Steindl et al., 2015). There is, however, no systematic analysis of whether commitment requests can backfire due to reactance. We close this gap and examine if commitment requests that contain behavioral restrictions lead individuals to choose deliberately the type of behavior that the request marks as (socially) undesirable.

It is also of note that the specific form of cheating studied in our field experiment (i.e., plagiarizing answers from neighbors) relates our work to research on peer effects and social interactions. Because we can better highlight the links to this methodological literature after we have introduced our field-experimental design, we relegate this discussion to Subsection 4.4. Before presenting the results from the field experiment, we will next discuss the laboratory evidence.

3 Commitment in the Basic Cheating Game

3.1 Conceptual Framework

To further structure the discussion of potential effects of commitments, we build on the simple conceptual framework for understanding cheating and lying behavior originally suggested by Kajackaite and Gneezy (2017). The framework can be easily adapted to nest all the previously discussed explanations for negative and positive commitment effects. In the following, we describe the utility function of an agent who faces a binary cheating decision, and also discuss how this function relates to the experimental design and our hypotheses.

Suppose an agent faces a binary decision to cheat or not. She observes the state of nature \( t \) and then self-reports the state. The agent has two option. She can either report the true state \( t \) or report a false state \( t' \). The monetary payoff from stating \( t \) is \( m_t \) and from reporting \( t' \) is \( m_{t'} \). This results in a monetary benefit of cheating of \( m_{t'} - m_t > 0 \). With \( p(m_{t'}, m_t) \) denoting the perceived probability of punishment and \( s(m_{t'}, m_t) \) denoting the perceived sanction in case of detection, we capture the

---

10 For instance, consumers for whom phosphate-containing detergents were banned reported more positive attitudes towards these products (Mazis et al., 1973), signs prohibiting the spraying of graffiti resulted in a more densely graffiti-covered wall (Pennebaker and Sanders, 1976), and threatening messages against littering resulted in more littering behavior than those appealing to social norms (Reich and Robertson, 1979). Vrugt (1992) shows that, after a policy that prohibited discrimination, negative attitudes against women were more pronounced, and Faith et al. (2004) reports that children responded to food restrictions by eating more of the restricted variety.
extrinsic cost of cheating by the expected sanction \( S[p(m_t', m_t), s(m_t', m_t)] \). Comparing only the monetary payoff and the extrinsic cost of cheating, the agent will cheat whenever \( m_t' - m_t > S[p(m_t', m_t), s(m_t', m_t)] \). This inequality illustrates the fundamental trade-off from Becker’s (1968) model on the economics of crime: An agent cheats if the benefits of dishonesty outweigh the expected costs.

As previously discussed, the agent’s decision may additionally depend on her intrinsic disutility of cheating. For example, a person might have a bad conscience if she realizes that she did not comply with her moral standards. We capture the disutility from not reporting truthfully by adding an intrinsic (psychological) cost of cheating \( 0 \leq C_i \leq \infty \) to the agent’s decision problem. Following Kajackaite and Gneezy (2017), we make the simplifying assumption that \( C_i \) is a fixed cost (i.e., it does not depend on the extent of cheating denoted by \( t' - t \) and \( m_t' - m_t \)).

Finally, we extend the framework such that it incorporates psychological reactance. Assume the agent faces a situation in which an external request to report truthfully is activated, indicated by \( r = 1; r = 0 \) if such a request is not made. In the case of an external request, a reactant agent obtains an additional fixed intrinsic utility of cheating \( 0 \leq R_i \leq \infty \). As discussed in Section 2, reactance makes cheating more attractive and reflects the psychological benefit of regaining one’s freedom of choice by not reporting truthfully under a request to do so. Note that we allow for heterogeneity in \( C_i \) and \( R_i \). Putting the extrinsic and intrinsic costs and benefits of cheating together, the agent will not report truthfully if:

\[
m_t' - m_t - S[p(m_t', m_t), s(m_t', m_t)] - C_i + R_i \cdot 1\{r = 1\} > 0, \tag{1}
\]

where \( 1\{\cdot\} \) is an indicator function.

Equation (1) summarizes our discussion on the channels through which commitment requests can affect cheating. On the one hand, commitment requests may increase the intrinsic disutility of cheating \( C_i \). For example, they may direct the agent’s attention to her moral reference point, making it psychologically more costly to cheat (Mazar et al., 2008). Other explanations in line with higher cheating costs under commitment are the feeling of guilt from breaking a promise (Charness and Dufwenberg, 2006), a preference for keeping one’s word (Ellingsen and Johannesson, 2004), lying aversion (Gneezy, 2005), or reputational costs (Abeler et al., 2018). On the other hand, and in line with the theory of psychological reactance (Brehm, 1966, 2009), reactant agents derive additional intrinsic utility from

\[\text{As we discuss in Subsection 3.4, the specification with two separate variables, } C_i \text{ and } R_i, \text{ is supported by the data. In particular, the data suggest that the intrinsic cost of cheating is uncorrelated with the subjects’ reactance type.}\]
cheating $R_i$ if they are requested to commit to truthful reporting. Different forms of commitment requests can thus lead to more or less cheating, depending on how sharply $C_i$ and $R_i$ are shifted.\textsuperscript{12}

### 3.2 Basic Cheating Game

We implemented a simple cheating game ($N = 303$) (a) to examine whether we can replicate the standard result that commitment requests can increase honesty and (b) to test whether freedom-restricting forms of commitment requests backfire. Before we introduce our treatments, we first present the basic cheating game that follows the computerized experiment in Abeler et al. (2018).\textsuperscript{13}

The design was as follows: After subjects entered the laboratory, the experimenter informed them that the session consisted of two parts: a survey and a short experiment. In the first part, subjects received a payoff of €4 for answering a 15-minute survey on the German inheritance tax schedule. We added this part to the experiment for two reasons. First, by placing other elements before the cheating decision, we followed the standard experimental protocol in the literature (see, e.g, Fischbacher and Föllmi-Heusi, 2013; Kajackaite and Gneezy, 2017). Second, and more importantly, we included this survey to introduce our commitment requests more naturally and mitigate experimenter demand effects. In particular, the experimenter placed the form to be signed by the subjects at the subjects’ workplaces before they entered the laboratory and reminded them to sign the “declaration concerning the behavioral rules in the laboratory” right at the beginning of the session. We hence connected the commitment to the entire session rather than to the cheating experiment.

At the beginning of the session’s second part, the participants read instructions from the computer screen (see the Appendix). The instructions informed subjects that the experiment would start with a computerized random draw of a number between one and six that they would have to self-report. Subjects also learned from the instructions that their additional payoff (i.e., the payoff in addition to the fixed payment for participating in the survey) would be €5 if they reported a 5 and zero

\textsuperscript{12}Conditional on the setting and the specific form of the declaration of compliance, commitment requests might also change the expected sanction $S[·]$. We discuss this issue when introducing the laboratory and the field-experimental designs.

\textsuperscript{13}We would like to thank Abeler et al. (2018) for providing code to replicate the computerized draw in their experiment. The experiment took place between May and December 2017 in the Laboratory for Experimental Research, Nuremberg. The Sessions lasted 45 minutes, including time for the subjects’ payment. The average participants earned €11.6, including the show-up fee. We programmed the experiment with z-Tree (Fischbacher, 2007) and recruited subjects with ORSEE (Greiner, 2015).
if they reported a number from the set \{0, 1, 2, 3, 4, 6\}.

The computerized random draw simulated the process of drawing a chip from an envelope. Subjects first saw an envelope containing six chips numbered between one and six on their screen (see the Appendix for screenshots). Subjects then clicked a button to start the draw. The chips were shuffled for a few seconds, and one randomly selected chip fell out of the envelope. On the next screen, subjects were asked to report their draw by entering the number into a field on the screen. After subjects had reported their number, the experimenter called the participants by their computer number and paid them anonymously for all parts of the session.

The fact that we computerized the random draw makes cheating observable at an individual level. This design element comes with the benefit of a much higher statistical power compared to approaches that identify cheating by evaluating the empirical distribution of self-reports against the expected distribution under truthful reporting (see, e.g., Fischbacher and Föllmi-Heusi, 2013). Nevertheless, if individuals believed that the instructions correctly described the experimental conditions, the expected (immediate) monetary sanction should nevertheless be zero. That is because we neither included a monetary punishment for cheating nor communicated a positive probability of such a punishment. Instead, the instructions highlighted that a subject’s payoff depended exclusively on the reported number.

We complemented the experimental data with survey data to elicit psychological reactance as a fundamental human trait. The survey-based standard measure for a subject’s reactance type, which we also use in this paper, is Hong’s Psychological Reactance Scale (Hong, 1992; De las Cuevas et al., 2014). The scale consists of 14 statements that approximate the degree to which a particular person shows reactance. For instance, one statement is “regulations trigger a sense of resistance in me,” and another one reads “when someone forces me to do something, I feel like doing the opposite.” To record the answers, we used a 5-point Likert Scale with higher (lower) values indicating stronger agreement (disagreement).

Importantly, to avoid spillovers between survey responses and behavior, we collected the survey data four days before the experiment. The procedure of survey data collection was as follows. Six days before a session, the subjects who had registered for the experiment received an invitation to take part in an online survey. The invitation email highlighted that subjects not answering the survey would be excluded from the upcoming session. Participants got 48 hours to answer the questionnaire, and we reminded subjects who had not completed the survey a few

\[^{14}\text{Before reporting their draw, subjects could also click a button to show the instructions and the payoff structure again. They could also click a button to display the result of the random draw again.}\]

\[^{15}\text{See the Appendix for full list of statements.}\]
hours before the deadline to participate. Answering the online survey took about five minutes. Participants received a fixed payoff of €2 for taking part.16

3.3 Commitment Treatments

We use the basic cheating game as a control condition. The treatment conditions differed from the control condition in that subjects signed a declaration right after they entered the laboratory and took their seats. The paper with the declaration displayed a short preamble highlighting that experiments at the Laboratory for Experimental Research Nuremberg are subject to certain behavioral standards and/or rules. Below the preamble, the paper included a brief declaration. As described in the following, we experimentally varied the content of the declaration and evaluated the impact of this variation on cheating behavior. One of the declarations followed the literature and aimed at evoking the self. Others had a more restrictive character and varied the degree to which reactance should be activated. A common feature of the different declarations was that, by signature, subjects expressed commitment that they would behave in a certain way during the experiment.

Our first treatment condition was ethics. This treatment followed the idea that an effective no-cheating declaration has two central properties. On the one hand, the treatment makes ethics salient (Mazar et al., 2008; Shu et al., 2011) or increases cheating costs through one of the other previously discussed channels (Ellingsen and Johannesson, 2004; Gneezy, 2005; Charness and Dufwenberg, 2006). On the other hand, it is not communicating behavioral restrictions that could trigger reactance. In this vein, in the ethics treatment, the declaration that followed the preamble highlighted the general principle of ethically sound behavior without expressing explicitly that certain types of actions were prohibited. It read: “I hereby acknowledge the principles of ethically sound behavior.” In line with the previous literature on commitment effects and reactance, our first hypothesis is:

**Hypothesis 1.** Signing a non-restricting declaration to behave ethically sound decreases cheating.

---

16 We informed the subjects that their survey responses would be linked with the experimental data using an individual key. 8.6% of the invited students did not complete the survey and hence did not participate in the experiment. No-shows were balanced across treatment groups.

17 The preamble read “The Laboratory for Economic Research Nuremberg (LERN) adheres to the ethical standards that were defined, e.g., by the German Research Foundation. One of the principles of ethically sound behavior is that data and findings must not be falsified. Today’s experiment is subject to the stated standards.”
In the second treatment condition, called NEUTRAL, the declaration that followed the preamble\(^{18}\) read: “I hereby declare that I will not violate the rules described in the instructions.” In contrast to the ETHICS condition, the declaration is neutral in the sense that it did not refer to any ethically loaded norm. We, therefore, expect this treatment to be a weaker self-identity prime. At the same time, the declaration contained an explicit restriction and requested a commitment to behave according to a given set of rules. In particular, given that the instructions required subjects to “enter the drawn number into the field provided for this purpose”, the treatment explicitly asked subjects to report a specific number honestly. Given the restrictive nature of this declaration, we expect the NEUTRAL treatment to be a stronger trigger for psychological reactance. The second hypothesis is:

**Hypothesis 2.** Signing a declaration to behave in accordance with an ethically neutral rule will not decrease cheating.

The third treatment condition was SANCTION.\(^{19}\) Subjects signed a declaration that read: “I hereby declare that I will not violate the rules described in the instructions. Violating the rules can lead to exclusion from future experiments.” Hence, this declaration included one additional sentence relative to the NEUTRAL treatment, highlighting the potential sanction in the case of non-compliance. We designed the declaration regarding the fact that reactance is believed to increase in the severity of the threat (Pennebaker and Sanders, 1976; Brehm and Brehm, 1981; Miron and Brehm, 2006; Steindl et al., 2015). That is because, by increasing the threat level, it becomes more salient that the freedom to choose freely is at risk. According to the theory of psychological reactance, harsher threats also make people feel uncomfortable, hostile, and angry, motivating them to exhibit the restricted behavior or force the threatening party to remove the threat (Brehm, 1966; Brehm and Brehm, 1981; Dillard and Shen, 2005; Rains, 2013).

Note that one potential side-effect of such a declaration is that it might have also altered the subjects’ expected sanction associated with a false report of the drawn number. However, as long as the perceived sanction increased, this works against the identification of a reactance effect.\(^{20}\) In summary, we expect the commitment

---

\(^{18}\) The preamble read “At the Laboratory for Economic Research Nuremberg (LERN), subjects participating in experiments have to adhere to certain rules. One of the rules requires subjects to follow the behavioral guidelines provided in the instructions for the experiment. Please sign the following declaration referring to this rule.”

\(^{19}\) The preamble was identical to the one in the NEUTRAL treatment.

\(^{20}\) As we have argued before, subjects trusting the instructions should not have expected any sanction.
in the SANCTION treatment to trigger a stronger reactance response. Our third hypothesis, therefore, is:

**Hypothesis 3.** Under the threat of being sanctioned, signing a declaration to behave in accordance with an ethically neutral rule will increase cheating.

### 3.4 Results for the Basic Cheating Game

We discuss the results of the laboratory experiment in two steps. First, we report how our treatments affect cheating. Second, we discuss the treatment-effect heterogeneity in the subjects’ reactance type.

**Average Treatment Effects**  Figure 1 displays the percent of cheaters per treatment group. The figure is based on all the subjects who did not draw a five and hence were able to cheat. In the CONTROL group, 27.7 percent of those subjects cheated by falsely reporting a five. Because we did not implement any form of commitment in the CONTROL group, this value approximates the share of cheaters in the absence of commitment effects. As can be seen from the figure, the subjects’ behavior in the treatment groups was substantially different from that in the CONTROL group. A comparison of the first two bars in the figure further shows that commitment to an ethically charged norm reduced cheating. This finding is in line with the experimental literature in psychology that has tested how signing an honor code and similar self-identity primes affect behavior. In our setting, the ETHICS treatment reduced the share of cheaters to 11.6 percent, a decrease of 16.1 percentage points (or 58 percent) relative to the control group (p-value = 0.032, t-test). Thus, our experimental data support Hypothesis 1, that signing a declaration to behave ethically sound decreases cheating.

Given that our experimental procedure allows us to reproduce the results of Mazar et al. (2008), Shu et al. (2011), and others qualitatively, the remaining treatment differences displayed in Figure 1 are quite striking. Turning to the NEUTRAL commitment, the share of cheaters was 36.5 percent, more than three times the share in the ETHICS conditions. The share of cheaters who signed a NEUTRAL commitment is not statistically different compared to the CONTROL group, however (p-value = 0.269, t-test). We thus find support for Hypothesis 2, that signing a declaration to behave in accordance with an ethically neutral rule does not decrease cheating.

---

21 The share of participants who drew a five was equal to 16.5 percent and balanced across treatments.
ing. Finally, the rightmost bar in Figure 1 demonstrates that the share of cheaters in the SANCTION treatment was 46.5 percent, i.e., 18.8 percentage points (or 67.9 percent) higher than in the CONTROL group (p-value = 0.023, t-test). This finding is in line with Hypothesis 3, stating that under the threat of being sanctioned, signing a neutral no-cheating declaration increases cheating. We conclude that the content of a commitment request matters: Although subjects in all the treatment groups express their commitment by signing a declaration of compliance, the effects differ. While an ethically charged form of commitment decreases cheating, commitment to an ethically neutral norm triggers more cheating.

The reported findings are robust to including a number of controls, such as age, gender, familiarity with similar games, and personal experience with similar games. Regressing an indicator for cheating on treatment indicators and control variables, the effect of ETHICS relative to CONTROL stays unaffected (effect size = 0.161; p-value = 0.036). The effect of NEUTRAL increases slightly from 0.088 (without controls) to 0.094 (with controls), but it remains statistically insignificant (p-value = 0.234). The impact of SANCTION increases from 0.188 to 0.230 and becomes even more significant (p-value < 0.01).

**Heterogeneity in Individuals’ Reactance Type**  This study is the first to document an adverse effect of commitment on agents’ honesty. Given the previous evidence that reactance is a crucial psychological force shaping individuals’ choice behavior, reactance seems like a plausible candidate to explain the backfiring effect. Our strategy to gather evidence on whether psychological reactance is a relevant channel through which commitment affects cheating behavior is to measure a subject’s general reactance type and to evaluate the effects of our treatments conditional on the magnitude of this personal trait. If reactance explains the individual’s cheating behavior, we expect the backfiring effect of a commitment to be strongest among types that are particularly prone to “pushing back” if their freedom of choice is restricted.

Figure 2 presents the findings. The Panels A1 to A4 show for the CONTROL and each of the treatment groups how the subjects’ cheating behavior related to their reactance score (higher scores indicate more reactant types). To calculate the score, one could also think of different explanations. One prominent alternative is motivational crowding out along the lines of Deci (1971), which can be interpreted as a decrease in an individual’s intrinsic cost of cheating $C_i$. Under motivational crowding, an extrinsic motivator (e.g., the declaration) undermines an individual’s intrinsic motivation to behave honestly. In terms of our conceptual framework, the psychological cost of cheating is reduced, possibly because an individual feels that her intrinsic motivation to report honestly is not acknowledged. Compared to psychological reactance, individuals do not perceive that their freedom (to cheat) is restricted or that a policy aims at restricting it.
we followed the factor analysis of De las Cuevas et al. (2014) and averaged over the subjects’ answers to eight of the fourteen statements.\textsuperscript{23} The resulting index approximates to what degree individuals engage in a restricted behavior to restore their behavioral freedom. We hence focus on reactant behavior, which is, given that the laboratory experiment captured cheating behavior, the relevant dimension. Each panel contains the results of the following linear probability model that regresses an indicator for lying $L_i$ on a third-order polynomial of the reactance score $\phi_i$:\textsuperscript{24}

$$L_i = \sum_{j=0}^{3} \beta_j \cdot (\phi_i)^j + u_i. \quad (2)$$

The graphs in Panels A1 to A4 display the empirical fraction of cheaters (bubbles) and the fraction of cheaters predicted by the model (red lines) for each realized score value. It also shows the 95% confidence bands.

We note that the models fit the data well and that there are obvious patterns. Three observations stand out. First, the graph for the control group (Panel A1) shows that in the absence of a commitment, the cheating behavior did not correlate with the reactance score: Subjects with high reactance scores showed a cheating behavior that was similar to the cheating behavior of subjects with low scores. Hence, to the extent that differences in the cheating behavior between subjects reflect heterogeneity in the intrinsic cost of cheating $C_i$, Panel A1 suggests that the intrinsic cheating costs were uncorrelated with the subjects’ reactance type. This supports the previously made theoretical assumption that the intrinsic utility of cheating due to reactance (captured by $R_i$) and the intrinsic cheating costs (captured by $C_i$) are independent terms in the utility function (see Subsection 3.1). Second, Panel A2 demonstrates that the cheating behavior under the ethically charged commitment was also unrelated to the subjects’ reactance type. Third, in both the NEUTRAL and SANCTION treatment, we observe a markedly different pattern pointing to a positive correlation between lying and the subjects’ reactance score (Panels A3 and A4). We first find indications that, under commitment to a neutral norm, reactance did matter for truthful reporting. Also note that in the SANCTION treatment, the share of cheaters among subjects at the top of the distribution of the reactance score was close to 100 percent.

\textsuperscript{23}The Appendix includes a list of all statements and indicates which statements we used to calculate the reactance score. Figure A1 in the Appendix also provides a summary of the main results for an index that uses all statements. This robustness check confirms our findings, but we note that the fit of the conditional expectation function to the data is worse. This suggests that by using all statements, we essentially add noise.

\textsuperscript{24}Changing the order of the polynomial does not affect our results qualitatively. For example, Figure A2 in the Appendix presents the results for a second-order polynomial regression.
The Panels B1 to B3 in Figure 2 show whether the reported treatment-specific heterogeneity also translates into heterogeneous treatment effects. To this end, we compare each of the treatments to the control group separately. The corresponding linear probability model is:

\[ L_i = \sum_{j=0}^{3} \beta_j \cdot (\phi_i)^j + \sum_{j=0}^{3} \gamma_j \cdot (\phi_i)^j \times T_i + u_i, \]  

(3)

where \( T_i \) is an indicator for the ETHICS, the NEUTRAL, or the SANCTION treatment.²⁵

Panel B1 shows the impact of ETHICS as a function of the reactance score. We do not find any discernible heterogeneity in the treatment effect. Thus, how commitment to an ethically charged norm affects a subject’s cheating behavior is not systematically related to the extent to which subjects are generally prone to “push back” behavioral restrictions. In contrast, the Panels B2 and B3 document a very systematic form of heterogeneity. The NEUTRAL treatment (Panel B2) did not affect the cheating behavior of subjects with low or moderately high reactance scores; however, for subjects with very high reactance levels, a neutral commitment sharply increases the lying probability. At the top of the reactance score distribution, our point estimate of the treatment effect is in the range of 50 percentage points. Turning to the impact of the SANCTION treatment (Panel B3), we find a similar but even more pronounced pattern. The effects for the most reactant types are also more extreme: At the top of the reactance score distribution, a neutral commitment that points subjects to a sanction increases the lying probability by more than 50 percentage points. Given that the average share of cheaters was less than 30 percent in the control group, this is a substantial upward shift in subjects’ cheating probability.

Taken together, the Figures 1 and 2 establish the following set of findings. First, using the ETHICS treatment, we replicate the result of previous studies that an ethically charged commitment can reduce cheating significantly. Second, extending the existing evidence, we show that, on average, commitment to an ethically neutral norm does not reduce cheating. Third, combining a commitment to a neutral norm with an additional trigger for reactance results in more cheating relative to the control group with no commitment. Fourth, the adverse effect of a neutral commitment on truthful reporting is driven by the subgroup of subjects with high reactance scores. While the evidence from our laboratory experiment does not provide ult-

²⁵The treatment-effect heterogeneity would show a causal mechanism (working via reactance) if the direct effect of commitment on cheating (i.e., the impact not running via reactance) was independent of a subject’s reactance type. See the causal mechanism literature for a discussion of this assumption (e.g., Baron and Kenny, 1986; Imai et al., 2011). Because this assumption is untestable, we interpret our results as suggestive evidence.
mate proof for a causal mechanism, the patterns in the data strongly suggest that the differences in how subjects responded to the treatments are in fact due to psychological reactance.

4 Adverse Effects of Commitment in the Field

The central takeaway message from our laboratory experiment is that requesting agents to commit to a no-cheating rule can backfire. This section demonstrates that this finding is not an artifact of the laboratory setting. For that purpose, we exploit a field experiment that manipulates whether or not students sign a no-cheating declaration before a university exam and then tests the effect of this intervention on the students' cheating behavior.

As we will demonstrate subsequently, the primary benefit of the exam setting is that it allows us to measure dishonesty objectively, which is notoriously difficult and often impossible outside the laboratory. This feature enables a detailed evaluation of how commitment affects cheating behavior. On the downside, the major drawback of the exam context is that pinning down the exact mechanism through which commitment affects cheating behavior is more difficult. In particular, we were unable to measure students' psychological reactance scores in the exam context; therefore, we cannot analyze the heterogeneity of the treatment response in the same way as we did in the laboratory. Furthermore, because the laboratory and the field settings are different in many dimensions (e.g., with respect to the perceived detection probability and the stakes at play), the mechanisms in the field experiment could be different from those in the laboratory. Our purpose in this subsection is to test the external validity of the result that commitment can backfire rather than to pin down the channels through which the students' behavior is affected precisely.

The remainder of this section is structured as follows. The Subsections 4.1 to 4.3 discuss the details of our field-experimental design. The field evidence is presented subsequently. Before evaluating the treatment effects, Subsection 4.4 considers each treatment separately and tests for the prevalence of cheating; it also includes a wide range of evidence on the structure of cheating. In particular, the analysis reveals that students mainly copy answers from direct neighbors sitting in the same row and that the overall level of cheating observed in the data is mostly due to low-performing students. Both results guide our analysis in Subsection 4.5, in which we estimate the effects of commitment.
4.1 Setting of the Field Experiment

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. 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. Likely, a first element that was a driver of cheating behavior is the perceived punishment in case of detection. The general rules are clear-cut: According to the exam regulations in the department, any detected attempts to deceive (such as copying answers from neighbors, using a mobile phone, etc.) lead to failure of the exam. It is also part of the exam regulations that supervisors in exams announce standardized examination rules by reading them aloud (see 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 will 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.

A second essential element affecting cheating behavior is the monitoring level, as it drives the detection probability in case of cheating attempts. Importantly, the

---

26 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.
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 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 4.4, 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 (see, e.g., McCabe et al. 2001 for a discussion of honor codes as a tool to counteract fraud in exams). Furthermore, in the years before the experiment, the department did not use any form of commitment requests to prevent cheating in exams.

### 4.2 Treatment Groups

The main purpose of the field experiment is to test how commitment affects cheating in exams. To that end, we randomly allocated students from two strata (gender and grade of university admission 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 also randomly assigned students to seats within the lecture halls and made sure that they took their preassigned seats.\(^{27}\)

The only difference between the **CONTROL** group and the **COMMITMENT** treatment was that students in the **COMMITMENT** treatment signed a declaration of com-

\(^{27}\)We informed students before the exam in which lecture hall they would be seated. Upon going there, 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.
pliance with the no-cheating rule. We placed this declaration on the first page of the exam materials (see Appendix for details). It read: “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.” The declaration was printed below a form in which students in all treatments had to fill in their names and university IDs. By choosing such a salient location, we aimed at directing the students’ attention to the declaration immediately before the beginning of the exam. Students were given extra time to read and complete the form and sign the declaration.

To further our understanding of the nature of this commitment request, several aspects of how we implemented commitment are worthy of note. First, by letting students sign the declaration, we changed the degree of commitment to an existing no-cheating rule relative to the 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. Furthermore, the announcement that students found cheating would be sanctioned was made immediately before the students got time to complete the first page of their exam materials. Hence, the type of commitment in the respective treatment was quite similar to the SANCTION treatment in the laboratory experiment.

As mentioned in the introduction, we also implemented a treatment with close monitoring of students (but no commitment). In this paper, we use the MONITORING treatment to substantiate that our methods can identify cheating. In particular, in the spirit of previous work highlighting that intensive monitoring can eliminate dishonesty (Kleven et al., 2011; Kleven, 2014; Pomeranz, 2015), we increased the monitoring intensity in the MONITORING treatment to a level that we expected would eliminate cheating. 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, these were as follows: In the MONITORING TREATMENT, we allocated additional supervisors to the rooms 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

\[28\] A post-exam check shows that all the students in the COMMITMENT treatment had signed the declaration of compliance.
and COMMITMENT). Importantly, in all rooms, supervisors remained at specific pre-defined spots throughout the exam. In CONTROL and COMMITMENT, supervisors took positions in the front of the room. In the MONITORING treatment, the spots where supervisors located were evenly distributed all-over the room. Figure A3 in the Appendix provides a stylized illustration of the room setups under baseline and under close monitoring.

4.3 Further Details of Design 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 rooms. 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 room 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 A1 in the Appendix demonstrates that all characteristics are balanced across the treatments.

Next, we discuss our sampling scheme. Figure 3 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 225 in the MONITORING treatment.

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

---

29Due to the overbooking procedure, some students could not be seated in their preassigned room. We relegated those students to additional halls that we excluded from the experiment.
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, we only implemented the control and the commitment treatments 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).  

4.4 Testing for the Prevalence and the 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 4.5 then describes how we can also exploit the method’s underlying idea to achieve our primary goal, 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 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 4 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 room. 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. It is apparent that many students who sat next to each other provided identical answers.

30 The reason for the remaining differences in the number of students in the control and commitment treatment in the second exam are differences in the capacity of the lecture halls.
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 hence 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 tackle 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 allows us to project how the similarity in actual neighbors’ answers would be 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 apparent that two types of correlations in neighbors’ answers that are not easily distinguishable from cheating 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, students may have faced the same institutional environments during the exams, leading to correlations in their answers. An example would be a lecture-hall effect in the sense that the room-specific examination conditions might have aligned the students’ answers in a room.

Having discussed the potential confounding factors, we can state the identify-
ing 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, 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 only used non-neighbors who either sat in the same lecture hall or even in the same row to construct counterfactual neighbor-pairs. This nets out lecture hall and row effects, respectively.

4.4.1 Do Students Cheat in Exams? Treatment-Specific Tests

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, the area of application of randomization testing is far-ranging. 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

---

32 The basic idea is as follows. Consider a linear regression \( Y = \alpha + \beta T + u \), where \( T \) is a binary treatment indicator. The randomization test for the hypothesis \( H_0: \beta = 0 \) proceeds as follows. First, we obtain an estimate of \( \beta \). Second, we randomly generate many placebo treatment assignments and estimate the associated regression coefficients. Third, we perform a hypothesis test by evaluating if the treatment effect is in the tails of the distribution of the placebo treatment effects.
another individual’s outcome (Athey et al. 2015; Athey and Imbens 2017). Furthermore, Falk and Ichino (2006) use an approach that is notably similar to ours to identify peer effects in co-workers’ productivities.

Testing Procedure Our paper firstly 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., creating counterfactual pairs of neighbors) and then recalculating what the similarity measure would have been if these assignments had been the real ones.

Our preferred parameterization of the randomization test is one that identifies plagiarism between direct neighbors who were sitting next to each other in a row and randomly reassigns students within rooms. Subsequently, we check the robustness of our results regarding these assumptions. For a student \(i\) in row \(r\) with a left neighbor \(i-1\) and a right neighbor \(i+1\), we can describe the testing procedure as follows:

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 rooms to seats (without replacement), and recompute the test statistic.

3. Repeat the second step \(M\) times, which generates a distribution of the test statistic with \(M\) values.

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, Figure 5 reports the results of our treatment-specific randomization tests \((M = 5000)\). Panel A shows the results for the control group and Panel B for the commitment treatment. In each panel, the vertical line depicts the average similarity in the answers of actual neighbors.
Δ. The bell-shaped curves represent the counterfactual distributions under the null hypothesis of no cheating.

Two findings emerge from Figure 5. 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 Δ is located far in the right tail of the distribution and we can, consequently, reject the null hypothesis of no above-normal similarity in the answers of actual neighbors (p-value = 0.001). This finding is the first 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). We will discuss the effect of commitment at length in Subsection 4.5, 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. If randomization tests indicate plagiarism, there should be consequently fewer or no statistically significant above-normal spatial correlations in the MONITORING treatment. The latter is precisely what we find: As documented in Figure A4 in the Appendix, we do not observe any evidence that the similarities in the MONITORING treatment are suspiciously high. The test statistic Δ is located in the center of the counterfactual distribution. As a result, we cannot reject the null hypothesis of no above-normal similarity in the answers of actual neighbors under close monitoring (p-value 0.999). This finding reinforces our confidence in the randomization test as a method to identify cheating.

Spatial Structure of Cheating and Level of Randomization The randomization tests reported in Figure 5 assume that students only copied answers from neighbors in the same row and exploit random assignment of students to seats within the lecture halls. As a robustness check, Figure 6 presents evidence for a variety of alternative specifications. The first type of robustness check examines whether or not students copied answers from other students sitting farther away. Consider the stylized seating plan in the lower right corner of Figure 6. We note that a student (yellow circle) may also have copied answers from fellow students sitting in the front or in the back (seat numbers 3 and 13), from students seated diagonally in

33 Throughout the paper, we use a total of four different neighbor definitions to test for cheating. To guard against spurious findings from multiple testing, we use a conservative Bonferroni adjustment to correct the reported p-values.
the front or in the back (seat numbers 2, 4, 12, and 14), or even second-order neighbors in the same row (seat numbers 6 and 10). If such cheating occurred, it goes undetected by the specification of the randomization tests reported in Figure 5. Our test statistics would be biased downwards.

Panel A in Figure 6 demonstrates that, in our context, copying from other students was confined to direct neighbors within rows. For comparison, Panel A1 replicates the evidence from Figure 5. The remaining panels report the results for second-order neighbors in the same row (Panel A2), neighbors sitting directly in the front and the back (Panel A3), and diagonal neighbors in the front and the back (Panel A4). Each of the figures considers the control group and the commitment treatment separately. It also contains the value of the relevant test statistic (red circle), the average value of the test statistic in the counterfactual distribution (blue circle), and the 95% confidence band for the counterfactual distributions (blue spikes). Considering the Panels A2 to A4, we cannot reject the null hypothesis of no above-normal similarity for any of the treatment groups. We conclude that students cheated in the CONTROL and COMMITMENT treatment groups, but we find measurable traces of copying from neighbors only for direct neighbors in a row.

Panel A in Figure 6 refers to specifications of the randomization test that randomly relocate students to seats within lecture halls. We prefer this more conservative resampling scheme because it decreases the probability of false positives by controlling for potential room effects. However, the flipside of this restriction is that it potentially increases the likelihood of false negatives. That is because, in this case, 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. Our second type of robustness check is to exploit such a randomization scheme. Panel B reports the respective results and demonstrates that our findings turn out to be robust.35

4.4.2 Students’ Ability and Cheating

The previous subsection demonstrated that actual neighbors shared a suspiciously high number of similar answers under baseline monitoring. In line with plagiarism, we expect to find especially strong correlations in neighbors’ responses if at least

34 Our results remain unchanged if we examine copying answers only from neighbors sitting in the front (as opposed to sitting both in the front and the back). For details, see Figure A5 in the Appendix.
35 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.
one of the two students is of low ability and was, therefore, less likely to know the solutions to the multiple choice problems herself. In the following, we demonstrate that it is indeed the low-ability students who cheat. However, cheating is confined to cases in which both students are of low ability. This, in turn, suggests that in our context, the similarity in incorrect answers is the most informative measure of cheating.

**Testing Procedure** Subsequently, we suggest a simple approach to study how the cheating behavior of students depends on ability. As randomization tests do not allow us to perform such an analysis, we employ a regression-based approach.

Let us start with a baseline model that does not include any additional regressors. The model, again, 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, the regression of interest is:

\[
Y_{mp} = \beta_0 + \beta_p \cdot N_p + u_{mp},
\]

(4)

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 average university admission grade (labeled A-level grade). The average A-level grade reflects a student’s performance in the final years in high-school and should, therefore, be a reasonable proxy for ability. Using A-level 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:

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

\[
= \sum_{i=1}^l \beta_{1i} \cdot B_{i,p} + \sum_{j=1}^l \beta_{2j} \cdot W_{j,p} + \sum_{i=1}^l \sum_{j \geq i} \beta_{3ij} \cdot B_{i,p} \times W_{j,p} + r_p.
\]

(5)

In this equation, \( B_{i,p} \) reflects the A-level 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 A-level grade was A, B, C, or D. Equivalently, $W_{i,p}$ are indicators for the A-level 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 (5) into (4). To complete the specification, we also add the non-interacted terms $\sum_{i=1}^{I} \alpha_{1i} \cdot B_{i,p}$, $\sum_{j=1}^{J} \alpha_{2j} \cdot W_{j,p}$, and $\sum_{i=1}^{I} \sum_{j=1, j \geq i}^{I} \alpha_{3ij} \cdot B_{i,p} \times W_{j,p}$. If $N_p$ is randomized, the OLS estimators of $\beta_{1i}$, $\beta_{2j}$ and $\beta_{3ij}$ are unbiased and consistent. The OLS estimators of $\alpha_{1i}$, $\alpha_{2j}$, and $\alpha_{3ij}$ pick up potential correlations between the grade variables and the error term $u_{mp}$. We consider observations from both exams and all rooms with baseline monitoring (CONTROL and COMMITMENT), and the regressions 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 7 presents the primary results of our linear probability models that either use identical incorrect answers (Panel A) or identical correct answers (Panel B) as the outcome variable. To construct this figure, we estimate model (4) and subsequently calculate the ANEs (5) for all potential grade combinations. The Panels A1 and B1 show the results. The horizontal axis depicts the ability (measured
by the A-level 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. 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: Lower ability students cheated, but cheating was confined to pairs in which both students were of low ability. To clarify and elaborate on this conclusion, let us consider Figure 7 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 A-level 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 an 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 (A-level 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 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 7 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-

36Importantly, 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.
ability students copy from each other, they share an above-normal number of identical incorrect answers. To the contrary, 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, focus on this measure when evaluating the effects of our treatments.

4.5 The Effect of Commitment on Cheating

Building on the evidence presented previously, this subsection examines our primary topic and tests whether a neutral commitment request also backfires in the field.

**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 (5), the following regressions account for treatment-specific neighbor effects. The decomposition of the neighbor effect becomes:

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

where \( C_p \) denotes an indicator for the COMMITMENT treatment. As before, we plug (6) into (4) and add \( \beta_3 \cdot C_p \) to the model. 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 A2 in the Appendix presents the results for regressions that instead rely on all types of identical answers (correct and incorrect).  

**Results** To explore the role of commitment for cheating, Table 1 reports the results of our linear probability models. We provide \( p \)-values in brackets.  

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

where \( C_p \) denotes an indicator for the COMMITMENT treatment. As before, we plug (6) into (4) and add \( \beta_3 \cdot C_p \) to the model. 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 A2 in the Appendix presents the results for regressions that instead rely on all types of identical answers (correct and incorrect).  

To that end, we note that the average probability of a given multiple choice question being answered correctly was 70%. Assuming that students who did not know

---

As described previously, identical correct answers do not contain much useful variation for measuring cheating. Adding identical correct answers to our outcome, hence, introduces noise to the dependent variable and makes the identification of the neighbor effect more difficult. However, if we increase the efficiency of our estimates by adding covariates to the model, the regression coefficients remain unchanged and become significant.

The number of observations derives 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).

---

37

38
the correct answer independently and randomly picked one of the four possible options, we predict the average probability of two students giving the same incorrect answer to be $3 \cdot (3/40)^2 = 0.0169$. This value is not a too bad approximation: In the CONTROL group, the baseline probability that two students that formed a pair of counterfactual neighbors shared an identical incorrect answer was 0.0364. The probability in the control group is, hence, not very different from a scenario in which students chose their answers randomly. Against this backdrop, we evaluate the coefficients in Table 1.

Beginning with the unconditional estimates in Column (1), we make three points of note. 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. Importantly, 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 qualitatively.

Finally, we turn to the main result of the field experiment and evaluate the coefficient of the interaction term Commitment $\times$ 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 the extent that mirrors the baseline effect of sitting next to each other.

A further question that naturally arises from the previous analysis is whether the increase in cheating in the COMMITMENT treatment did lead to better grades. The evidence suggests that this was not the case. Regressing the percentage of answers solved correctly on a dummy for the COMMITMENT treatment, we consistently find, across a number of specifications, that the coefficient of the treatment indicator
is very small and insignificant. This result is entirely in line with the previously reported evidence that predominantly low-ability pairs of students copied primarily incorrect answers from each other.

**Further Robustness Checks** The estimations reported in Table 1, 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 A-level 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 A3 displays the results of regressions that include an indicator variable for each room to control for room-specific differences in the similarity of the students’ answers. The findings are virtually unchanged. Table A4 estimates regressions at the room level. The estimation sample consists of (a) all pairs of actual neighbors in the same row and (b) all pairs of counterfactual neighbors who were sitting in the same room (as opposed to counterfactual neighbors who were sitting in the same row). We also cluster the standard errors at the room level. The point estimates of the coefficient for Actual Neighbors remains unchanged, the estimates of the ones for Commitment × Actual Neighbors slightly increases, and the corresponding p-values for both types of coefficients are always below 0.077. The results are also robust against estimating logit models instead of linear probability models (see Table A5 in the Appendix). Table A6 pools the data across all treatments (CONTROL, COMMITMENT, and MONITORING) and reports esti-

---

39 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 control group answered 72.4% of all multiple choice questions correctly. We obtain very similar results if we additionally cluster on the individual level or run separate regressions in both exams.

40 Prompted by the article of 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.
mations that also include a neighbor effect for the MONITORING treatment. The coefficient of the interaction term Monitoring × Actual Neighbors is negative and statistically significant, and we cannot reject the hypothesis that the similarity in actual neighbors’ answers under close monitoring was not different from 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.

The central message to take away from the analysis of the exam data is that even in a natural field context, neutral commitment requests can have the unintended consequence of inducing more cheating relative to a setting without commitment. Thus, the results from the field experiment confirm our main finding from the laboratory. A natural interpretation in line with psychological reactance is that less talented students value the freedom to decide to cheat in exams or enjoy cheating more if plagiarism is explicitly forbidden. A restricting commitment request would then make cheating a more attractive option and, in our context, trigger more plagiarism. In a similar vein, less talented students could operate under the perception of a social norm that copying answers from neighbors in exams is legitimate. The commitment request could then lead to more cheating in defense of the behavioral freedom to act following the existing social norm.

5 Conclusion

Many public institutions, including public schools and universities and the administrations of the tax, transfer, and judiciary system, require individuals and firms to commit to truthful reporting by signing no-cheating declarations. Although those institutions commonly implement commitment requests of all kinds, little is known about whether (and why) they affect the reporting behavior of agents. In particular, the existing literature on the effects of commitment has almost exclusively focused on morally charged declarations of compliance as a particular form of requesting commitment.

In this paper, we extend the literature in three important dimensions. First, we use a laboratory experiment to test how different commitment requests affect truthful reporting. The studied requests range from a morally charged declaration of compliance to a morally neutral declaration highlighting restrictions on the agent’s behavioral freedom. The evidence from our laboratory experiment demonstrates

\[ \beta_p = \beta_1 + \beta_2 \cdot C_p + \beta_3 \cdot M_p + r_p, \]

where \( M_p \) is an indicator for the MONITORING treatment.

---

34
that these commitment requests have vastly diverging effects. While a commitment to the morally charged declaration decreases misreporting, a commitment to the neutral declaration that threatens to punish backfires. One potential explanation for the backfiring effect is psychological reactance (Brehm, 1966), stating that individuals value a situation with a high degree of behavioral freedoms and tend to push back if they feel those freedoms are restricted. In the case of a request to commit to a specific rule, this line of reasoning suggests that subjects perceive the act of non-compliance with the rule as an option to restore their behavioral freedom.

In this vein, the second addition to the literature is that we present evidence of reactance as a channel through which commitment affects cheating. We use survey data collected before the experiment and demonstrate that the backfiring effect of a commitment request that highlights the loss of behavioral freedom is driven by subjects who we classify as highly psychological reactant types. While this piece of evidence is not an ultimate test of causality, the patterns in the data strongly suggest that reactance is indeed the driving force behind the backfiring effect of commitment.

Our third contribution to the literature is to provide field-experimental evidence demonstrating that the adverse effect of commitment requests extends to an important public institution operating in the real world. The field experiment exploits the setting of a university exam and implements a control group and a commitment treatment in which students are requested to sign a morally neutral declaration of compliance with a no-cheating rule. Our main finding is that students in the commitment treatment plagiarize significantly more relative to students in the control group. We conclude that the adverse effect of commitment requests on truthful reporting is a real-world phenomenon and not just an artifact of the laboratory setting.

The evidence presented in this paper speaks to contexts in which individuals and firms are routinely requested to commit to no-cheating rules. Our findings from the laboratory and the field show that commitment requests can have very beneficial but also quite damaging effects on honest behavior. Whether public institutions can induce honesty by requesting commitment depends crucially on a thoughtful implementation. Most importantly, they should carefully avoid triggering the perception that the requested commitment implies a loss of behavioral freedom.

References


Table 1: Responses to Commitment: 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>Commitment</td>
<td>0.0008 [0.7997]</td>
</tr>
<tr>
<td>Actual Neighbors</td>
<td>0.0073*** [0.0012]</td>
</tr>
<tr>
<td>Commitment × Actual Neighbors</td>
<td>0.0088** [0.0190]</td>
</tr>
<tr>
<td>Multiple Choice FE</td>
<td>No</td>
</tr>
<tr>
<td>Pair Controls</td>
<td>No</td>
</tr>
<tr>
<td>Share of Identical Incorrect Answers amongst Counterfactual Pairs in Control Group</td>
<td>0.0364</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 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 A-level 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.
**Figure 1: Cheating Behavior by Treatment**

Notes: This figure shows the percent of individuals who cheat in the ETHICS treatment ($N = 43$), the CONTROL group ($N = 65$), the NEUTRAL treatment ($N = 74$), and the SANCTION treatment ($N = 71$). We focus on individuals who did not draw a five and hence could cheat. The figure also includes 95% confidence intervals for linear probability models with robust standard errors.
Figure 2: Cheating Behavior and Psychological Reactance

A: Treatment-Specific Cheating Behavior


B: Heterogeneous Treatment Effects

B1: Ethics vs Control  B2: Neutral vs Control  B2: Sanction vs Control

Notes: This figure shows how cheating behavior relates to subjects’ reactance score. Panels A1-A4 show the empirical fraction of cheaters (bubbles) and the predicted fraction of cheaters (red lines) by group, conditional on the reactance score. The underlying linear probability model is $L_i = \sum_{j=0}^{3} \beta_j \cdot (\phi_i)^j + u_i$, where $L_i$ is an indicator for lying, and $\phi_i$ is a subject’s reactance score. The Panels B1-B3 display heterogenous treatment effects (relative to the control group) for the ETHICS, the NEUTRAL, and the SANCTION treatment, respectively. The bubbles represent empirical differences between treatments, and the red lines indicate the treatment effects obtained from the linear probability model $L_i = \sum_{j=0}^{3} \beta_j \cdot (\phi_i)^j + \sum_{j=0}^{3} \gamma_j \cdot (\phi_i)^j \times T_i + u_i$, where $T_i$ is an indicator for the respective treatment. The spikes indicate 95% confidence bands (Huber-White standard errors).
Figure 3: Overview of Field-Experimental Design

Exam 1

Sample
1007 students sampled
766 students took exam

Control
432 students sampled
333 students took exam

Commitment
265 students sampled
208 students took exam

Monitoring
310 students sampled
225 students took exam

Exam 2

Control
262 students sampled
204 students took exam

Commitment
170 students sampled
149 students took exam

Grading / Data Collection

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 4: 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 room. 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 5: Randomization Tests: Cheating by Treatment Group

A: Control

B: Commitment

Notes: This figure shows the results for treatment-specific randomization tests. The null hypothesis is: students do no cheat. The tests (a) identify cheating in the form of plagiarism, (b) focus on cheating to the left and right, and (c) calculate the treatment-specific test statistic as $\Delta = \frac{1}{N} \sum_{i} s_{i, i-1} + s_{i, i+1}$. $s_{i, i+1}$ reflects the share of all multiple-choice problems that $i$ and her left $i-1$ or right $i+1$ neighbor answer identically. Both panels use observations from both exams. In each panel, the vertical line represents the index value for the actual seating arrangement. The bell-shaped curve shows the counterfactual distribution of the test statistic on the basis of Epanechnikov kernels. The $p$-values are 0.001 (Control) and 0.001 (Commitment).
Figure 6: Spatial Structure of Cheating and Randomization Schemes

A: Randomization within Rooms

A1: Row
(Seats: 7,9)

A2: 2nd Order Row
(6,10)

A3: Column
(3,13)

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

B: Randomization within Treatments

B1: Row
(Seats: 7,9)

B2: 2nd Order Row
(6,10)

B3: Column
(3,13)

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

Stylized Seating Plan

Notes: This figure 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 empirical value of the relevant test statistic (red circles), the average value of the test statistic in the counterfactual distribution (blue circles), and the 95% confidence bands for the counterfactual distributions (blue spikes).
Figure 7: 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 (4) 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 5). 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. A-level 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.
Online Appendix
(not for publication)
Table A1: Balancing Checks Field Experiment

<table>
<thead>
<tr>
<th></th>
<th>Exam 1</th>
<th></th>
<th></th>
<th>Exam 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Commitment</td>
<td>Monitoring</td>
<td>Difference</td>
<td>Control</td>
<td>Commitment</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(6)</td>
<td>(7)</td>
</tr>
<tr>
<td>Gender (Female = 1)</td>
<td>0.54</td>
<td>0.56</td>
<td>0.50</td>
<td>0.02</td>
<td>-0.04</td>
<td>0.53</td>
</tr>
<tr>
<td>University Admission Grade</td>
<td>2.47</td>
<td>2.48</td>
<td>2.50</td>
<td>0.02</td>
<td>0.03</td>
<td>2.48</td>
</tr>
<tr>
<td>Math Proficiency</td>
<td>0.75</td>
<td>0.73</td>
<td>0.73</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.75</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>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td>Age</td>
<td>19.6</td>
<td>19.6</td>
<td>19.6</td>
<td>-0.04</td>
<td>-0.03</td>
<td>19.7</td>
</tr>
<tr>
<td>Foreign</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>Bavaria</td>
<td>0.81</td>
<td>0.83</td>
<td>0.84</td>
<td>0.02</td>
<td>0.02</td>
<td>0.80</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 heteroscedasticity-robust 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. Female is a dummy variable (female = 1). University Admission Grade is the overall grade of the university admission qualification (from High School), 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. Foreign and Bavaria are dummies indicating where students obtained their university admission (reference group: German states other than Bavaria). The sample consists of the 766 students in the experiment. Gender and University Admission Grade were used for stratification.
Table A2: Responses to Commitment: All Identical Answers

<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>-0.0029</td>
<td>-0.0029</td>
<td>-0.0001</td>
<td>-0.0001</td>
</tr>
<tr>
<td>Actual Neighbors</td>
<td>0.0144***</td>
<td>0.0144***</td>
<td>0.0120**</td>
<td>0.0120**</td>
</tr>
<tr>
<td>Commitment × Actual Neighbors</td>
<td>0.0105</td>
<td>0.0105</td>
<td>0.0143*</td>
<td>0.0143*</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>

Share of Identical Answers amongst Counterfactual Pairs in Control Group 0.5745

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 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 A-level 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 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>
</tbody>
</table>

Share of Identical Incorrect Answers amongst Counterfactual Pairs in Control Group

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<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 incorrect answers. Estimates are based on linear probability models. Column (1) includes room indicators to control for room 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 A-level 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: Responses to Commitment: Regressions at Room-Level

<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>0.0004</td>
<td>0.0004</td>
<td>−0.0004</td>
<td>−0.0004</td>
</tr>
<tr>
<td></td>
<td>[0.8408]</td>
<td>[0.8404]</td>
<td>[0.8450]</td>
<td>[0.8454]</td>
</tr>
<tr>
<td>Actual Neighbors</td>
<td>0.0072**</td>
<td>0.0072**</td>
<td>0.0071**</td>
<td>0.0071**</td>
</tr>
<tr>
<td></td>
<td>[0.0225]</td>
<td>[0.0225]</td>
<td>[0.0188]</td>
<td>[0.0188]</td>
</tr>
<tr>
<td>Commitment × Actual Neighbors</td>
<td>0.0092*</td>
<td>0.0092*</td>
<td>0.0091*</td>
<td>0.0091*</td>
</tr>
<tr>
<td></td>
<td>[0.0778]</td>
<td>[0.0773]</td>
<td>[0.0575]</td>
<td>[0.0571]</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>

Share of Identical Incorrect Answers amongst Counterfactual Pairs in Control Group: 0.0367

Number of Clusters: 8
Number of Observations: 1,238,279

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. The estimation sample consists (a) of all pairs of actual neighbors and (b) all pairs of counterfactual neighbors who were sitting in the same room. 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 A-level 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 room level; p-values in brackets.
Table A5:
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</td>
<td>0.0241</td>
<td>0.0073</td>
<td>0.0081</td>
</tr>
<tr>
<td>Actual Neighbors</td>
<td>0.1896***</td>
<td>0.1977***</td>
<td>0.1697***</td>
<td>0.1792***</td>
</tr>
<tr>
<td>Commitment × Actual Neighbors</td>
<td>0.1848**</td>
<td>0.1953**</td>
<td>0.1705**</td>
<td>0.1823**</td>
</tr>
<tr>
<td>Multiple Choice FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</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 logit models. The table shows logit coefficients. 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 A-level 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 A6:
Responses to Commitment and Monitoring: Actual and Counterfactual Pairs

<table>
<thead>
<tr>
<th>Dependent Variable: Indicator for Identical Incorrect Answer</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>Monitoring</td>
<td>0.0033</td>
<td>0.0033</td>
<td>0.0036</td>
<td>0.0036</td>
</tr>
<tr>
<td></td>
<td>[0.4445]</td>
<td>[0.4443]</td>
<td>[0.3970]</td>
<td>[0.3965]</td>
</tr>
<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>[0.7989]</td>
<td>[0.7990]</td>
<td>[0.9524]</td>
<td>[0.9530]</td>
</tr>
<tr>
<td>Actual Neighbors</td>
<td>0.0073***</td>
<td>0.0073***</td>
<td>0.0067***</td>
<td>0.0067***</td>
</tr>
<tr>
<td></td>
<td>[0.0009]</td>
<td>[0.0009]</td>
<td>[0.0014]</td>
<td>[0.0014]</td>
</tr>
<tr>
<td>Monitoring × Actual Neighbors</td>
<td>−0.0095**</td>
<td>−0.0095**</td>
<td>−0.0092**</td>
<td>−0.0092**</td>
</tr>
<tr>
<td></td>
<td>[0.0240]</td>
<td>[0.0245]</td>
<td>[0.0406]</td>
<td>[0.0408]</td>
</tr>
<tr>
<td>Commitment × Actual Neighbors</td>
<td>0.0088**</td>
<td>0.0088**</td>
<td>0.0080**</td>
<td>0.0080**</td>
</tr>
<tr>
<td></td>
<td>[0.0176]</td>
<td>[0.0175]</td>
<td>[0.0124]</td>
<td>[0.0123]</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>

**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 A-level 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>
<thead>
<tr>
<th>Hall</th>
<th>Control</th>
<th>Hall</th>
<th>Commitment</th>
<th>Hall</th>
<th>Monitoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Students per Supervisor</td>
<td></td>
<td>Students per Supervisor</td>
<td></td>
<td>Students per Supervisor</td>
</tr>
<tr>
<td>1</td>
<td>51.3</td>
<td>5</td>
<td>56.5</td>
<td>9</td>
<td>9.2</td>
</tr>
<tr>
<td>2</td>
<td>49.8</td>
<td>6</td>
<td>47.5</td>
<td>10</td>
<td>8.5</td>
</tr>
<tr>
<td>3</td>
<td>38.0</td>
<td>7</td>
<td>44.5</td>
<td>11</td>
<td>8.0</td>
</tr>
<tr>
<td>4</td>
<td>29.0</td>
<td>8</td>
<td>30.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Treatment-specific Averages

46.4 46.6 8.4

Notes: This table contains information on the number of students per supervisors in each lecture hall. It also shows the weighted average of the monitoring intensity within the control, the commitment, and the monitoring treatment, respectively (weights: number of students in lecture hall).
Figure A1: Cheating Behavior and Reactance (Score Includes All Items)

A: Ethics vs Control

A1: Predicted Fraction of Cheaters  
A2: Treatment Effect

B: Neutral vs Control

B1: Predicted Fraction of Cheaters  
B2: Treatment Effect

C: Sanction vs Control

C1: Predicted Fraction of Cheaters  
C2: Treatment Effect

Notes: This figure shows how cheating relates to the subjects’ reactance score. We calculate the score by averaging over all items of the Hong’s Psychological Reactance Scale (Hong, 1992). Panel A compares ETHICS and CONTROL; Panel B contrasts NEUTRAL and CONTROL; Panel C compares SANCTION and CONTROL. The right-hand panels show heterogeneous treatment effects obtained from the linear probability model \( L_i = \sum_{j=0}^{3} \beta_j \cdot (\phi_i)^j + \sum_{j=0}^{3} \gamma_j \cdot (\phi_i)^j \times T_i + u_i \). \( L_i \) is an indicator for lying, \( \phi_i \) is a subject’s reactance score, and \( T_i \) is an indicator for the respective treatment. The left-hand panels depict the treatment-specific predictions of the same model. The models use Huber-White standard errors and the figure includes 95% confidence bands.
A: Treatment-Specific Cheating Behavior

A1: Control

A2: Ethics

A3: Neutral

A4: Sanction

B: Heterogeneous Treatment Effects

B1: Ethics vs Control

B2: Neutral vs Control

B2: Sanction vs Control

Notes: This figure shows how cheating behavior relates to the subjects’ reactance score. The Panels A1-A4 show the empirical fraction of cheaters (bubbles) and the predicted fraction of cheaters (red lines) by group, conditional on the reactance score. The underlying linear probability model is \( L_i = \sum_{j=0}^{2} \beta_j \cdot (\phi_i)^j + u_i \), where \( L_i \) is an indicator for lying, and \( \phi_i \) is a subject’s reactance score. The Panels B1-B3 display heterogeneous treatment effects (relative to the control group) for the ETHICS, the NEUTRAL, and the SANCTION treatment, respectively. The bubbles represent empirical differences between treatments, and the red lines indicate the treatment effects obtained from the linear probability model \( L_i = \sum_{j=0}^{2} \beta_j \cdot (\phi_i)^j + \sum_{j=0}^{2} \gamma_j \cdot (\phi_i)^j \times T_i + u_i \), where \( T_i \) is an indicator for the respective treatment. The models use Huber-White standard errors and the figure includes 95% confidence bands.
Figure A3: Monitoring Conditions in the Field Experiment

**Baseline Monitoring**

**Close Monitoring**

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 A4: Randomization Test: Cheating in the Monitoring Treatment

Notes: This figure shows the result for the randomization test in the MONITORING treatment. The null hypothesis is: students do no cheat. The test (a) identifies cheating in the form of plagiarism, (b) focuses on cheating to the left and right, and (c) calculates the test statistic as \( \Delta = \frac{1}{N} \sum_{i} s_{i,\pm 1} \). \( s_{i,\pm 1} \) reflects the share of all multiple-choice problems that \( i \) and her left \( i - 1 \) or right \( i + 1 \) neighbor answer identically. The vertical line represents the index value for the actual seating arrangement. The bell-shaped curve shows the counterfactual distribution of the test statistic on the basis of Epanechnikov kernels. The \( p \)-values is 0.999.
Figure A5: 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 (a) examines the spatial structure of cheating (Panel A) and (b) 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 empirical value of the relevant test statistic (red circles), the average value of the test statistic in the counterfactual distribution (blue circles), and the 95% confidence bands for the counterfactual distributions (blue spikes).
Thank you for participating in today’s experiment!

Please read the instructions carefully. For completing the online survey, you will receive 2 Euro. The show-up fee is 4 Euro. For answering the questionnaire, you will receive 5 Euro (first part of today’s session). There is a possibility to earn another 5 Euro in the following experiment (second part of today’s session).

Your entire earnings will be paid to you in cash at the end of the second part of the experiment. Also note that this is a computer-based experiment. The data will be analyzed anonymously.

---

Please read the instructions carefully. When you have finished reading the instructions, click the CONTINUE button.

You will then see six chips with the numbers 1, 2, 3, 4, 5, and 6. Click the START button. The chips will be placed in the envelope. The envelope will be shuffled a couple of times. Then one of the chips will be drawn randomly, and this particular chip will fall out of the envelope.

Please enter the number you have drawn into the field provided for this purpose. You will receive 0 Euro if you enter the numbers 1, 2, 3, 4, or 6. You will receive 5 Euro if you enter a 5. That is, if you fill in a 5, you will receive an additional payoff of 5 Euros; you will not receive any additional payoff if you fill in any other number.

Once you have entered your number, you will receive your payoff anonymously.
Six chips above envelope

Six chips in envelope
Chips are shuffled

One chip falls out of the envelope
Subjects report number
Psychological Reactance Scale (Hong, 1992)

The following statements concern your general attitudes. Read each statement and please indicate how much you agree or disagree with each statement. If you strongly agree, mark a 5. If you strongly disagree, mark a 1. If the statement is more or less true of you, find the number between 5 and 1 that best describes you. There are no right or wrong answers. Just answer as accurately as possible.

Behavioral and Cognitive Component (De las Cuevas et al., 2014)

1. Regulations trigger a sense of resistance in me.
2. I find contradicting others stimulating.
3. When something is prohibited, I usually think, “That’s exactly what I am going to do.”
4. I consider advice from others to be an intrusion.
5. Advice and recommendations usually induce me to do just the opposite.
6. I am content only when I am acting of my own free will.
7. I resist the attempts of others to influence me.
8. When someone forces me to do something, I feel like doing the opposite.

Affective Component (De las Cuevas et al., 2014)

9. The thought of being dependent on others aggravates me.
10. I become frustrated when I am unable to make free and independent decisions.
11. It irritates me when someone points out things, which are obvious to me.
12. I become angry when my freedom of choice is restricted.
13. It makes me angry when another person is held up as a role model for me to follow.
14. It disappoints me to see others submitting to standards and rules.
Front sheet of exam materials in the field experiment

Answer Sheet for Exam in „Principles of Economics“

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

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

Declaration

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

________________________
Signature
List of official announcements to be made before written exam

Announcements
Please read out loud before the exam starts!

1. Bags, folders, etc. need to be set aside such that you cannot access them during the exam.
2. Smoking is prohibited in the lecture hall.
3. Take care to provide legible handwriting. Unreadable parts will not be marked.
4. Cheating is forbidden and any attempt to deceive will lead to failure of the exam; i.e., your exam will be graded with a 5.0.
   Attempts to deceive are:
   (a) if you are not sitting in your assigned seat
   (b) if you communicate with your neighbors or copy answers from neighbors
   (c) if your cellphone is not switched off
   (d) if you possess or use unauthorized materials during the exam

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

   Now is your last chance to hand in unauthorized materials. There will be check-ups during the exams.

5. Please make sure that you received the correct exam materials. Stay in your seats until the exam has ended. The supervisors will collect your answers sheets after the exam. It is your responsibility to hand in the answer sheets.

6. The examination period starts after we have distributed the examination materials (i.e., the questions). Don’t touch the examination materials until the start of the exam was announced. Questions concerning the problem sets will not be answered.

7. If you become sick during the exam, you have to report this immediately. After the exam, you cannot claim that you were physically incapable of taking the test.

8. Please only use the provided pen to fill in the answer sheet. This facilitates the automated scanner-based evaluation of the multiple-choice answer sheets. Please make sure that the pen remains at your work desk after the end of the exam. We will collect the pens separately from the exam materials.

9. You now have 5 minutes time to complete the first page of the answer sheet. Instructions how to fill in the multiple-choice answer sheet are provided on the second page.
Additional Laboratory Experiments

Several complementary laboratory experiments were reported in earlier versions of the paper. For completeness, in the following, we document the experimental designs and the main findings.

Experiment on Attitudes Towards Cheating

We ran a laboratory experiment on attitudes towards cheating. The purpose of the experiment was to test if a commitment request can, in principle, shift the intrinsic psychological cost of cheating $C_i$. Table A8 summarizes the experimental design. The experiment implemented a control and commitment treatment. The sequence of events was as follows. In the commitment treatment, the experimenter asked participants to sign a declaration to adhere to a no-cheating rule and checked if all participants signed the declaration. Similar to the Neutral treatment, the declaration formulated an explicit behavioral restriction but neither referred to any ethically loaded norm nor included a reactance trigger (such as a threat). It read:

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

In the control treatment, the declaration-stage of the experiment was not implemented. In both treatments, the experimenter distributed a questionnaire on attitudes toward cheating (see the following pages), and participants answered it. Next, the experimenter distributed and read out paper-based instructions for a real-effort task. The experimenter then started the stage with the real-effort task. In this stage, participants had five minutes to answer 20 multiple choice questions on principles of economics. The multiple choice questions were similar to those in Exam 1 in the field experiment. After finishing the real-effort task, participants handed in their answers to the experimenter who counted the number of correct answers. Subjects were paid according to their performance, using a piece rate per correct answer of €0.50. The show-up fee was €5. The experimenter made sure that participants did not learn about other participants’ performance in the real effort task. The experiment took place in April 2014 in the Laboratory for Experimental Research Nuremberg. In total, 62 economics and business administration students participated in the experiment. The sessions lasted about 60 minutes. Participants earned €9.2, or $12.7 including the show-up fee. All individual characteristics were balanced across the treatments.
Question 1
Please indicate for each of the following actions whether you think it can always be justified, never be justified, or something in between. Please choose a value from the following list: 1 means “never justifiable”, 10 means “always justifiable”.

<table>
<thead>
<tr>
<th>Action</th>
<th>Never justifiable</th>
<th>Always justifiable</th>
<th>Don’t know</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using unauthorized materials during exams</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using unauthorized aid from other students during exams</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Giving unauthorized aid to other students during exams</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using unauthorized materials during laboratory experiments</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Question 2
Imagine you earn €5 during this laboratory experiment without using unauthorized materials. Assume that using unauthorized materials cannot be detected. Please indicate the minimum payoff under usage of unauthorized materials that would let you choose this option.

___________________________ €

Question 3
Imagine you obtain the mark 2.3 in the exam of a compulsory course. Assume that using unauthorized materials cannot be detected. Please indicate the minimum mark under usage of unauthorized materials that would let you choose this option.

____________________________

Question 4
Imagine you participate in an exam and using unauthorized materials cannot be detected. Please indicate to what extent you would use unauthorized materials and what kind of behavior you would expect from other students. Please choose a value from the following list for yourself and regarding other students: 1 means “no usage of unauthorized materials”, 10 means “full usage of unauthorized materials”.

<table>
<thead>
<tr>
<th>No usage of any unauthorized material</th>
<th>Full usage of unauthorized material</th>
<th>Don’t know</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9 10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

You .................................................................

Other students ....................................................
**Question 5**

For every of the statements below, please indicate to what extent you agree with them. Please choose a value from the following list: 1 means "I totally disagree", 4 means "I totally agree".

<table>
<thead>
<tr>
<th>Statement</th>
<th>Totally disagree</th>
<th>Totally agree</th>
<th>Don't know</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>I'm always trying to figure myself out</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generally, I'm not very aware of myself</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I reflect about myself a lot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I'm often the subject of my own fantasies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I never scrutinize myself</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I'm generally attentive to my inner feelings</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I'm constantly examining my motives</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I sometimes have the feeling that I'm off watching myself</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I'm alert to changes in my mood</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I'm aware of the way my mind works when I work through a problem</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Question 6**

a) Age

b) Gender

c) Begin of studies (year)

d) Field of Studies

e) Number of computer
Table A8: Laboratory Experiment on Attitudes Towards Cheating: Design

<table>
<thead>
<tr>
<th>Stage of Experiment</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Commitment</td>
</tr>
<tr>
<td></td>
<td>Subjects Sign Declaration</td>
</tr>
<tr>
<td>2</td>
<td>Questionnaire</td>
</tr>
<tr>
<td>3</td>
<td>Real Effort Task</td>
</tr>
<tr>
<td>4</td>
<td>Evaluation by Experimenter</td>
</tr>
</tbody>
</table>

Table A9 summarizes our main findings on attitudes toward cheating. Column (1) shows the effect of the commitment request on the acceptance of cheating. To elicit acceptance, the questionnaire asked subjects to indicate the degree to which they believe it is justifiable to cheat on exams, using a scale from 1 (never) to 10 (always).\textsuperscript{42} Responses show that subjects in the commitment treatment consider cheating significantly less acceptable than those in the control treatment. The effect is substantial, indicating a 26.5% decrease in acceptance relative to the control group ($p$-value 0.025, Mann-Whitney U). Column (2) reports the effect of commitment on the reported psychological cost of cheating. To measure psychological cost, we described a scenario in which compliant behavior in a laboratory experiment yields a payoff of €5. We then asked subjects about the minimal monetary payoff that would make them engage in cheating. We interpret the difference between the answer and the €5 payoff under compliance as the monetary payoff that is necessary to compensate individuals for the psychological cost of cheating. In the commitment treatment, the average psychological cost is €20.8, compared to €14.7 in the control group. This indicates a 41.4% increase in the reported psychological cost of cheating ($p$-value 0.086, Mann-Whitney U). We conclude that requesting a commitment to a no-cheating rule affects subjects mentally: Signing a no-cheating rule triggers a shift in subjects’ attitudes toward cheating and increases the psychological cost of rule violations.

Column (3) complements the evidence on shifts in attitudes by showing the effect of the commitment request when subjects predict their cheating behavior in a hypothetical exam situation. We do not find any difference between the treatment and control group. Hence, although requesting commitment affects subjects’ stated attitudes toward cheating, the subjects themselves do not predict that this would affect their own behavior in an exam situation. This finding is consistent with the

\textsuperscript{42}The wording of the questions on the acceptance of cheating is similar to that of the World Values Survey questions on the acceptance of tax evasion and free-riding.
Table A9: Effects of Commitment on Attitudes Towards Cheating

<table>
<thead>
<tr>
<th></th>
<th>Acceptance of Cheating (1)</th>
<th>Psychological Cheating Cost (2)</th>
<th>Predicted Own Cheating Behavior (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commitment vs. Control</td>
<td>-0.960** [0.025]</td>
<td>6.07* [0.086]</td>
<td>-0.233 [0.738]</td>
</tr>
<tr>
<td>Average Outcome in Control Group</td>
<td>3.62</td>
<td>14.7€</td>
<td>5.20</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>61</td>
<td>50</td>
<td>60</td>
</tr>
</tbody>
</table>

Notes: This table shows average treatment effects of commitment on attitudes towards cheating. Column (1) reports the treatment effect on how acceptable subjects consider cheating in university exams. Responses lie between one (do not accept at all) and 10 (fully accept). Column (2) shows the treatment effect on the psychological cost associated with cheating, measured as the minimum additional payoff required to make cheating in a laboratory experiment acceptable. Column (3) displays the treatment effect on subjects’ predicted own cheating behavior in a hypothetical exam situation, standardized to lie between one (would not take advantage of cheating opportunity) and 10 (would take full advantage). Differences in the number of observations between columns are due to different numbers of students with non-response or “don’t know” answers. Significance: ** 5%; * 10%, inference based on Mann-Whitney U-tests (p-values in brackets).

evidence on actual cheating behavior: In our main laboratory experiment, we show that ethically neutral declarations without explicit reactance triggers do not affect cheating behavior. However, we note the possibility that the significant effects on attitudes toward cheating in columns (1) and (2) are driven by experimenter demand effects.

Laboratory Experiment on Cheating Behavior with Real Effort Task

In an earlier version of this paper, we studied the impact of a no-cheating declaration on cheating behavior in an experiment that identified cheating indirectly. In revising this paper, we decided to switch to the cheating game proposed by Abeler et al. (2018) that allows us to observe cheating on an individual level. We changed to this setting to obtain a higher statistical power and to be able to control for individual characteristics. Note also that we did not vary the type of the no-cheating declaration as part of the old experiment. Instead, we analyzed the effect of a single declaration that formulated an explicit behavioral restriction but neither referred to any ethically loaded norm nor included a reactance trigger. In line with the NEUTRAL treatment that had the same properties, we find that this type of declaration was not affecting behavior. While the experimental design hence slightly differed from the cheating game discussed in Section 3, the results of the old experiment are nested in the previously described ones (comparison between NEUTRAL treatment and CONTROL group).
As for the experimental design, the old experiment was based on the design of Mazar et al. (2008) and identified the average effect of requesting a commitment to a no-cheating rule on the subjects’ self-reported performance in a real effort task. The experiment also determined the extent of cheating by including a treatment with perfect monitoring. The experiment hence implemented three treatments: control, commitment, and monitoring.43 Table A10 summarizes the experimental design.

Table A10: Laboratory Experiment on Cheating Behavior: Design

<table>
<thead>
<tr>
<th>Stage of Experiment</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Commitment</td>
</tr>
<tr>
<td>1</td>
<td>Subjects Sign Declaration</td>
</tr>
<tr>
<td>2</td>
<td>Real Effort Task</td>
</tr>
<tr>
<td>3</td>
<td>Self-Evaluation</td>
</tr>
</tbody>
</table>

The sequence of events in the experiment was as follows: The experimenter started the experiment by distributing materials (including a questionnaire) and reading out instructions.44 In the commitment treatment, the experimenter asked participants to sign a declaration containing an ethically neutral no-cheating rule and checked if all participants signed this declaration. The declaration read:

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

In the monitoring and control treatment, the declaration-stage of the experiment was not implemented. In all treatments, the experimenter then started the stage with the real-effort task. Participants had five minutes to answer 20 multiple choice questions on principles of economics. The multiple choice questions were similar to those in Exam 1 in the field experiment. To test if the results depended on the type of real effort task, we also re-ran the experiment using the same matrix task as Mazar et al. (2008). This task consisted of 20 matrices containing a set of 12 numbers with two decimal digits. Participants got four minutes to find two numbers per matrix adding up to 10.45

In the control and the commitment treatment, the experimenter distributed a solution sheet once the stage with the real-effort task was over. The participants were

43We implemented all treatments in separate sessions.
44Apart from questions on age, gender, the field of study and beginning of studies, the questionnaire includes a field for the number of correct answers.
45According to Mazar et al. (2008), subjects do not view this task as one reflecting math abilities or intelligence.
then asked to check how many of their answers were correct and to self-report this number. Furthermore, the experimenter did not recollect the answer sheets in the treatments with self-evaluation. Hence, subjects could overstate their performance in these treatments without facing any risk of being detected.\textsuperscript{46} In contrast, in the monitoring treatment, participants handed in their solutions to the experimenter who counted the number of correct answers. After the participants (control and commitment treatment) or the experimenter (monitoring treatment) counted the number of correct answers, the participants answered a questionnaire. In the control and the commitment treatment, participants self-reported the number of correct answers in a separate field on this questionnaire. In these two treatments, the experimenter then paid participants according to their self-reported performance, using a piece rate per correct answer of €0.50. The show-up fee was €5. In the monitoring treatment, subjects were paid according to their actual performance. The experimenter made sure that participants did not learn about the other participants’ performance and/or self-reports.

The experiment took place in December 2013 in the Laboratory for Experimental Research, Nuremberg. In total, 184 students participated in the experiments. 91 subjects worked on the multiple choice task, while 93 subjects worked on the matrix task. Sessions lasted 45 minutes on average. Participants on average earned €9.8, including the show-up fee. The samples were balanced across treatments in all observable characteristics.

Table A11 summarizes the results for the multiple choice and the matrix task. The outcome studied is the share of correct answer/tasks in percent. We report average treatment effects in relative terms (i.e., the average treatment effect divided by the mean outcome in control group). Column (1) shows the impact of monitoring (monitoring versus control treatment) focusing on the multiple choice task. Without any opportunity to cheat, the subjects’ performance is 37.5 percent lower compared to the condition with self-evaluation ($p$-value = 0.006). This demonstrates that subjects made heavy use of the available cheating opportunity. Column (2) shows the effect of commitment (commitment versus control treatment). The commitment request had no significant effect on the students’ performance under self-evaluation ($p$-value = 0.278). Columns (3) and (4) report the effects of monitoring and commitment from the experiment using the matrix task, confirming the results from the multiple choice task.

\textsuperscript{46}We made sure that participants realized the detection probability was zero by stating upfront that answer sheets would not be recollected. We also made sure that participants did not learn about other participants’ (actual or self-reported) performance.
Table A11:  
Effects of Monitoring and Commitment on Performance in the Lab

<table>
<thead>
<tr>
<th>Treatment vs. Control</th>
<th>Multiple Choice Task Effects in %</th>
<th>Matrix Task Effects in %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monitoring (1) Commitment (2)</td>
<td>Monitoring (3) Commitment (4)</td>
</tr>
<tr>
<td>Treatment vs. Control</td>
<td>-37.5*** [0.006]</td>
<td>-37.5*** [0.000]</td>
</tr>
<tr>
<td></td>
<td>13.5 [0.278]</td>
<td>11.3 [0.222]</td>
</tr>
<tr>
<td>Performance in Control Group</td>
<td>52.0%</td>
<td>57.1%</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>60</td>
<td>61</td>
</tr>
</tbody>
</table>

Notes: This table reports average treatment effects on the share of correct answers from real effort tasks in percent. Columns (1) and (2) show the effects of monitoring and commitment from the experiment using the multiple choice task. Columns (3) and (4) report the same effects from the experiment using the matrix task. Significance: *** 1%, based on Mann-Whitney U-tests (p-values in brackets).