Supplementary Online Appendix to

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

Tobias Cagala, Ulrich Glogowsky, Johannes Rincke

A Tables **2**

B Figures **9**

C Lower Bound for the Share of Cheating Pairs **25**

D Share of Cheaters: Alternative Test **29**

E Suggestive Evidence on Channels **30**

E.A Experimental Design 30

E.B Results 35

A Tables

Table A1: Average Neighbor Effect Under Baseline Monitoring

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

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

Table A2: Alternative Counterfactual: Neighbor Effect Under Baseline Monitoring

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

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

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

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

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

Table A4: Responses to the Treatments: Alternative Counterfactual

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

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

Table A5: Responses to the Treatments: All Identical Answers

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

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

Table A6: Monitoring Intensity by Lecture Hall

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

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

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

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

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

B Figures

Figure B1: List of Announcements

List of official announcements to be made before written exam

Announcements
Please read out loud before the exam starts!

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

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

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

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

Figure B2: Front Sheet and Honesty Declaration

Front sheet of exam materials in the field experiment

Answer Sheet for Exam in

„Principles of Economics“

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

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

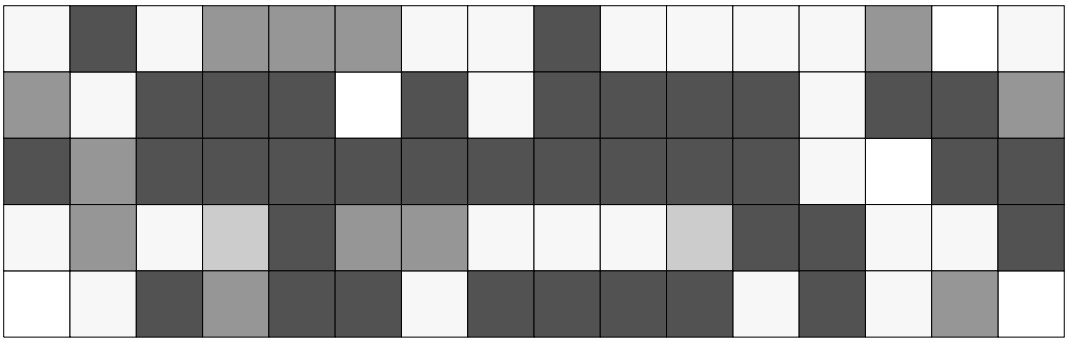
Declaration

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

_____

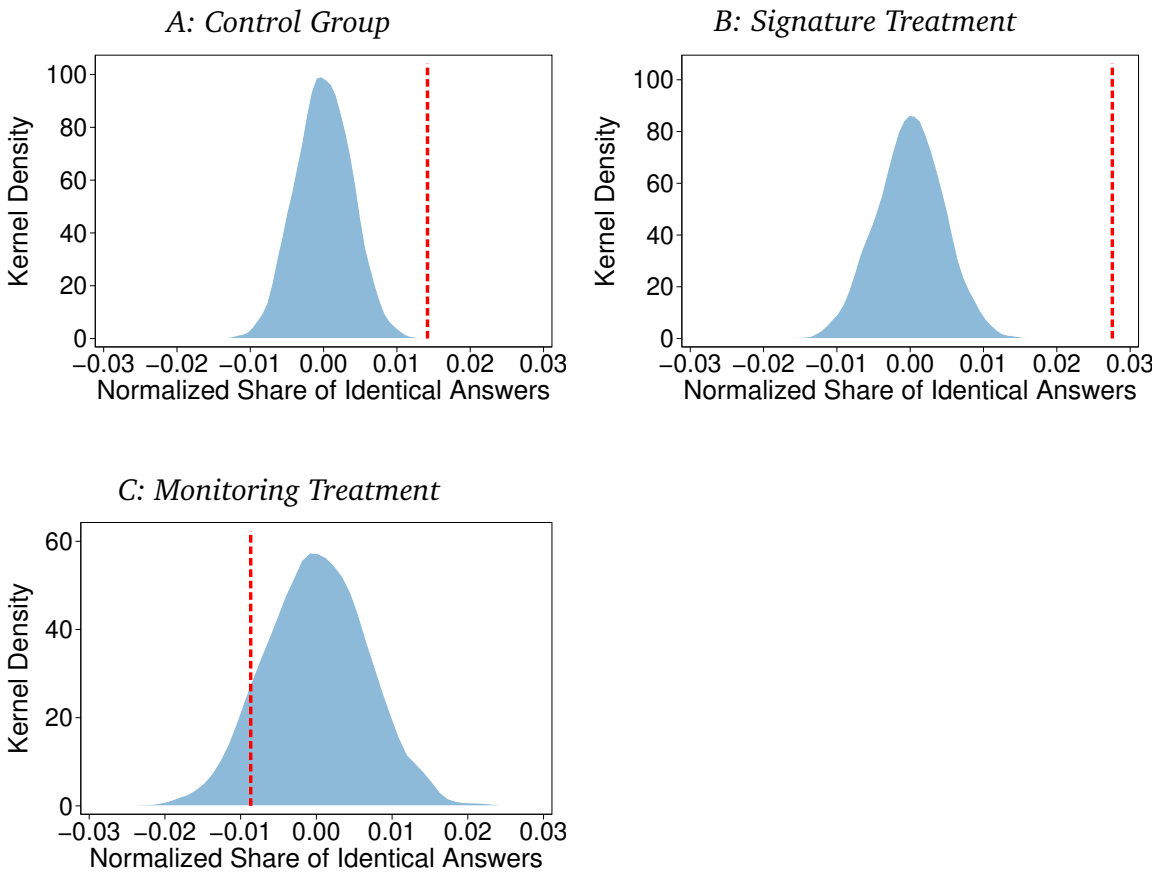
Signature

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



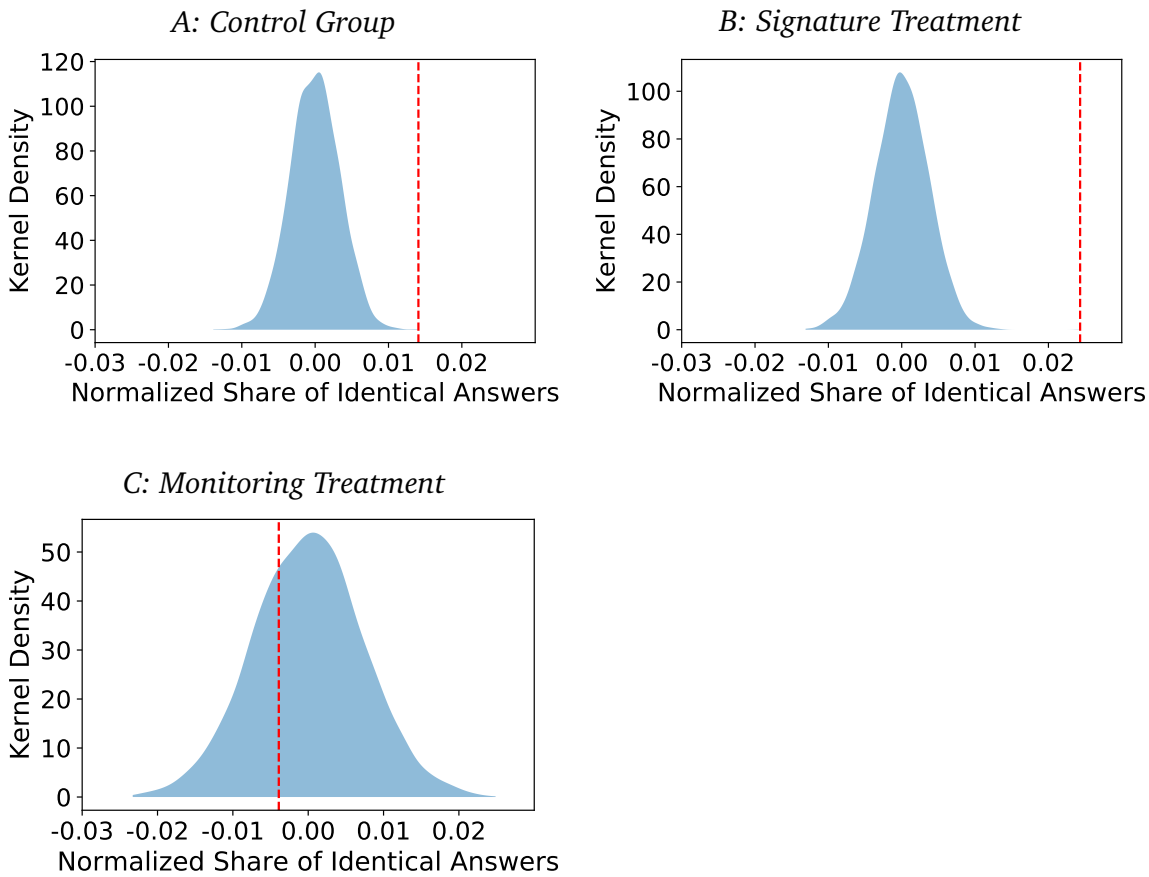
Notes: This figure provides an idea of what kind of data patterns our methods exploit. It visualizes the spatial pattern of answers to one multiple-choice problem in a selected control-group hall. Each rectangle represents a student, and the shade of the rectangle indicates the student's answer. Because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

Figure B4: Cheating by Treatment Group



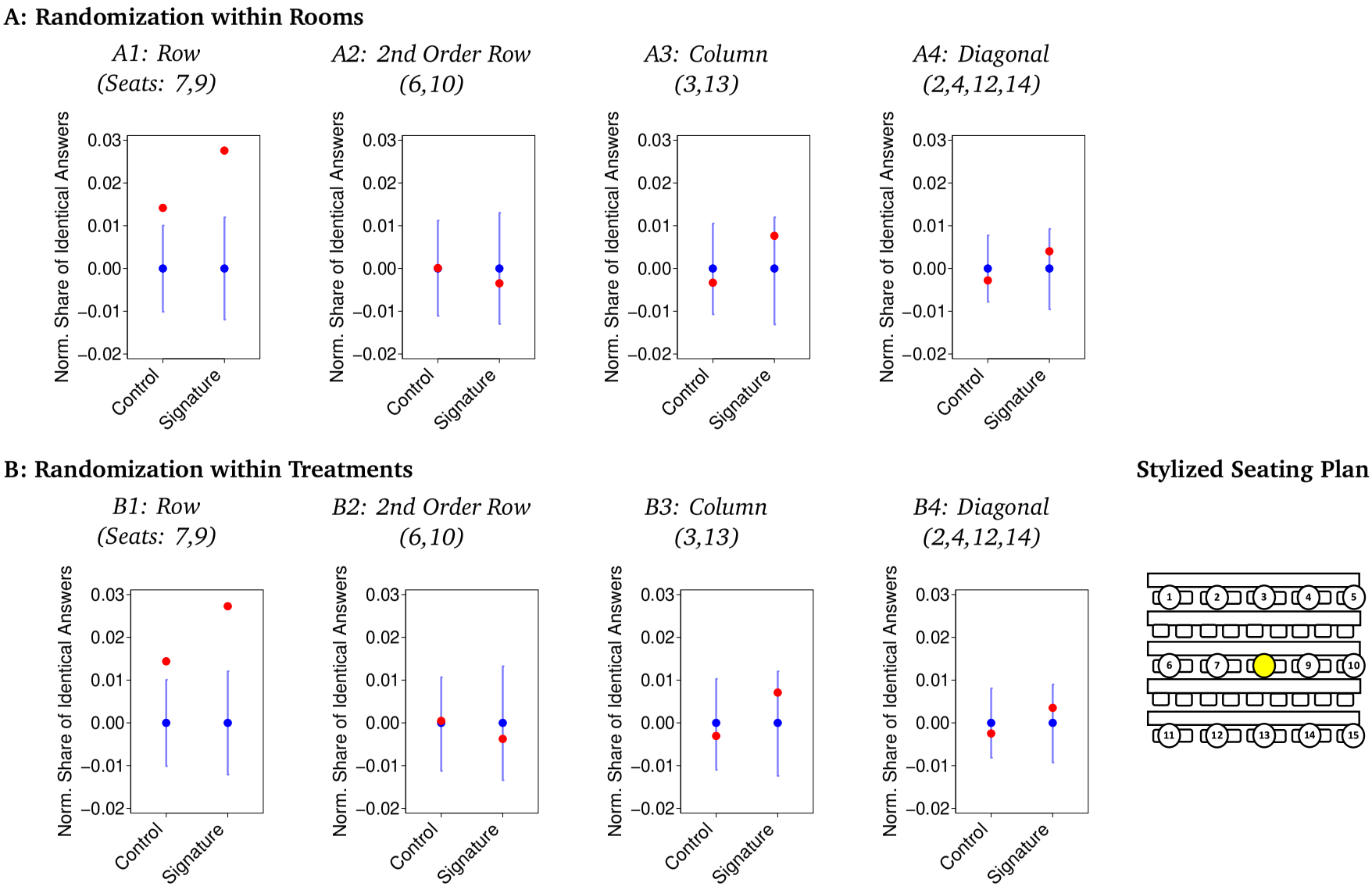
Notes: This figure shows the treatment-specific results for our randomization tests. Panel *A* focuses on the control group, Panel *B* on the signature treatment, and Panel *C* on the monitoring treatment. The vertical lines represent the test statistic derived from the actual seating arrangement. The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

Figure B5: Alternative Counterfactual: Cheating by Treatment Group



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

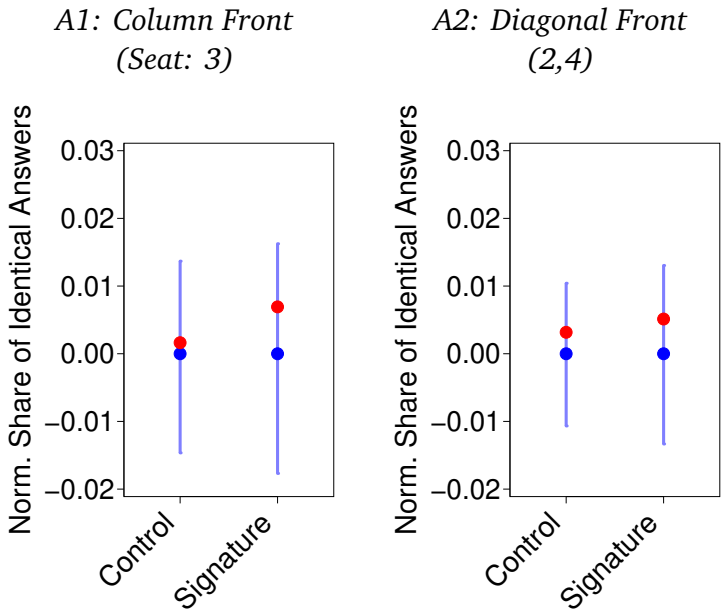
Figure B6: Spatial Structure of Cheating and Randomization Schemes



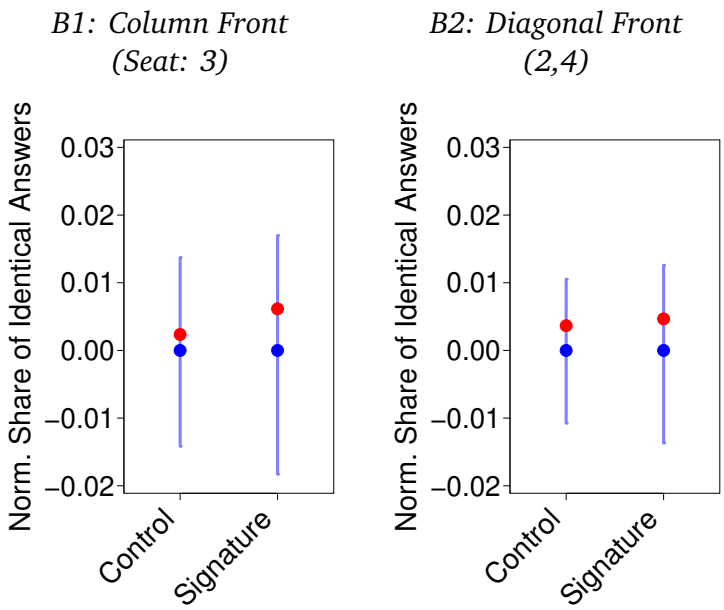
Notes: This figure serves two purposes. First, it tests the robustness of our results to the randomization schemes. Panel *A* randomizes individuals within rooms. Panel *B*, instead, resamples individuals within treatments (i.e., also across halls). Both reassignment procedures ensure that the counterfactual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

Figure B7: Spatial Structure of Cheating and Randomization Schemes

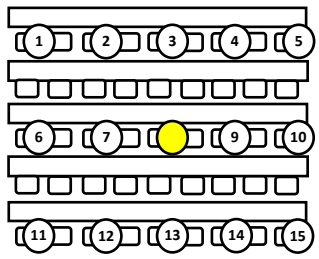
A: Randomization within Rooms



B: Randomization within Treatments

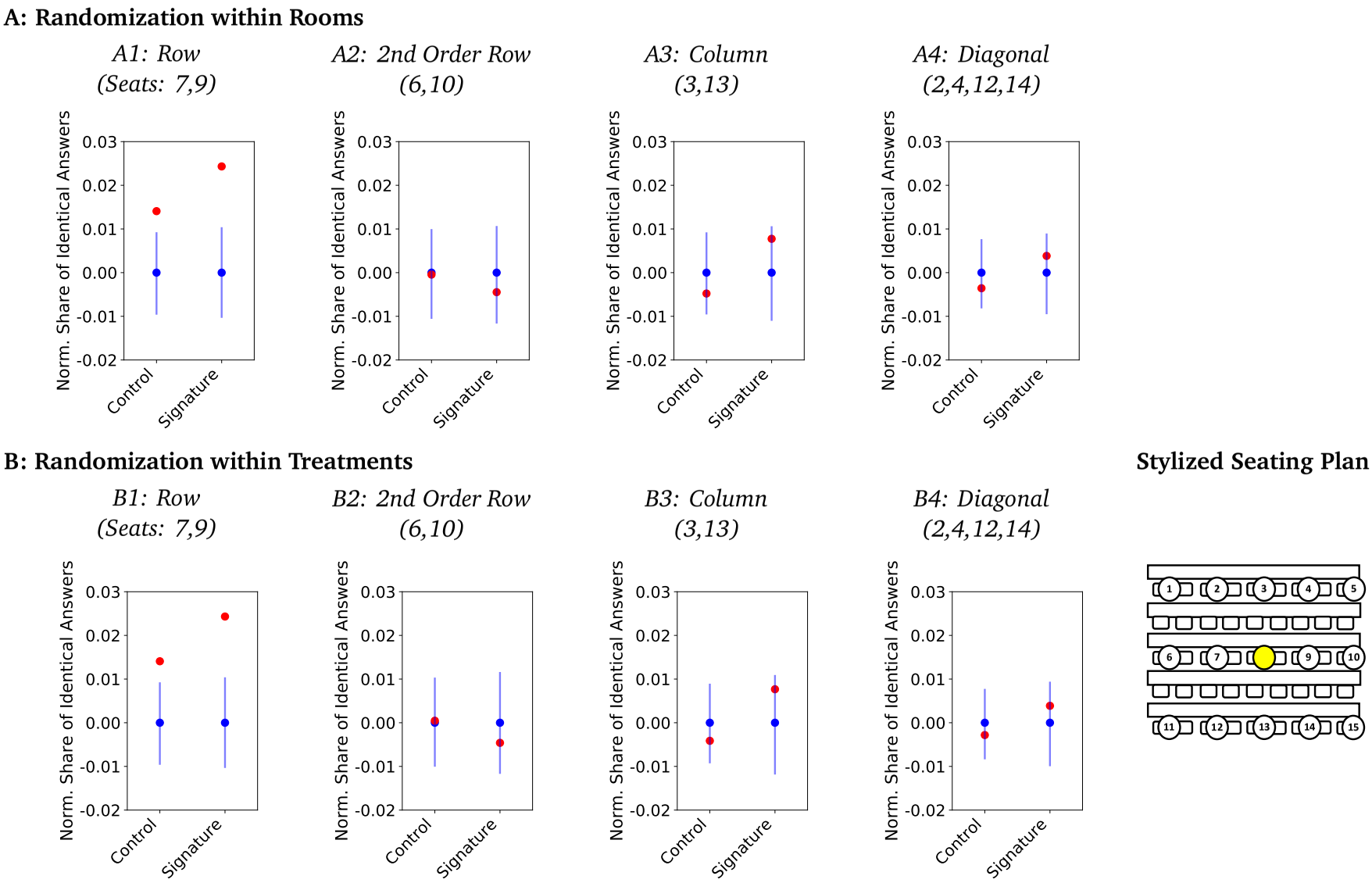


Stylized Seating Plan



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

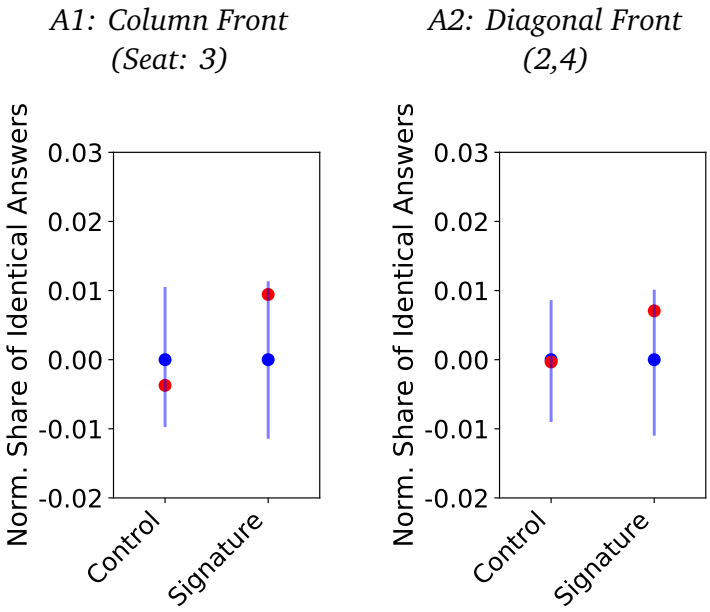
Figure B8: Alternative Counterfactual: Spatial Structure of Cheating and Randomization Schemes



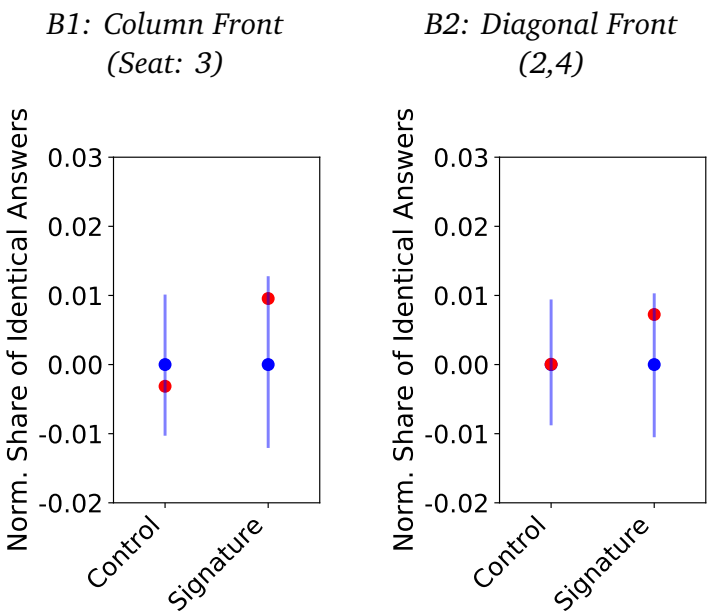
Notes: This figure serves two purposes. First, it tests the robustness of our results to the randomization schemes. Panel *A* randomizes individuals within rooms. Panel *B*, instead, resamples individuals within treatments (i.e., also across halls). Now, counterfactual pairs can consist of two students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

Figure B9: Alternative Counterfactual: Spatial Structure and Randomization Schemes

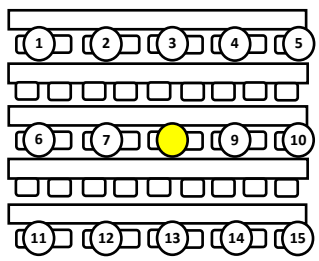
A: Randomization within Rooms



B: Randomization within Treatments

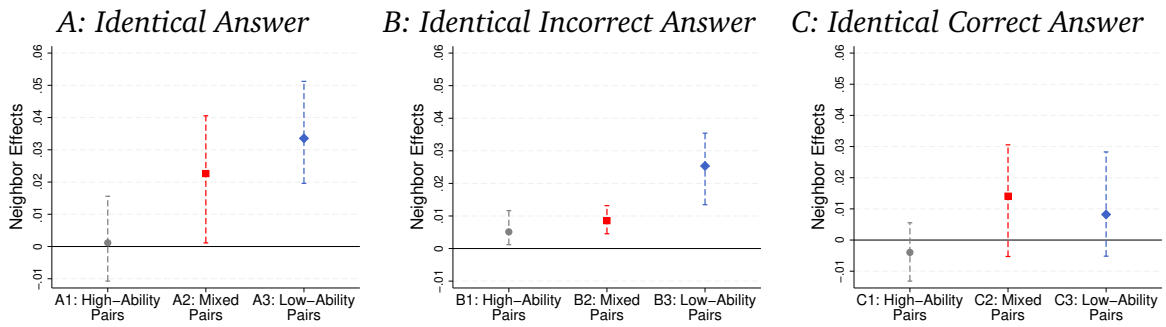


Stylized Seating Plan



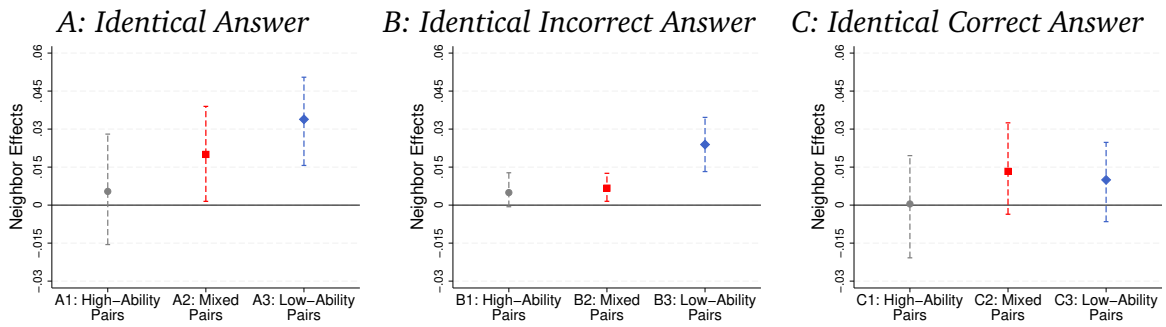
Notes: This figure serves two purposes. First, it tests the robustness of our results to the randomization schemes. Panel *A* randomizes individuals within rooms. Panel *B*, instead, resamples individuals within treatments (i.e., also across halls). Now, counterfactual pairs can consist of two students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

Figure B10: Grade Heterogeneity: Estimation with Control Variables



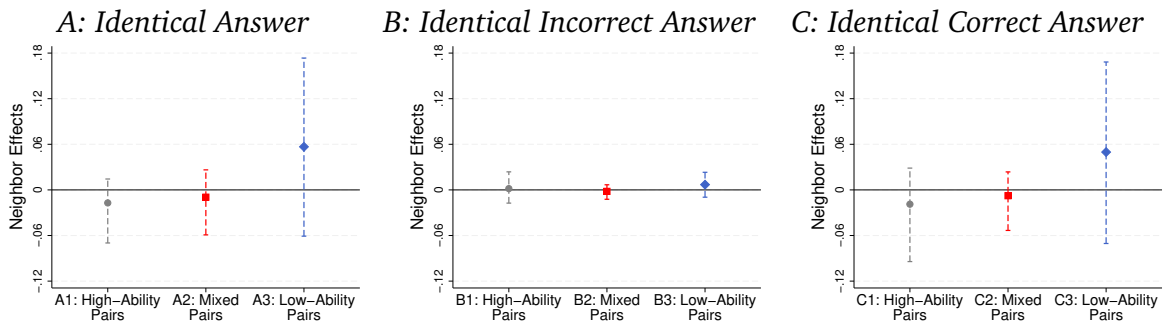
Notes: This figure examines how the students' ability (proxied by high-school performance) relates to their cheating behavior (under baseline monitoring). To construct this figure, we use model (2) to estimate the effect of being a pair of actual neighbors on the probability that two students give identical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

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



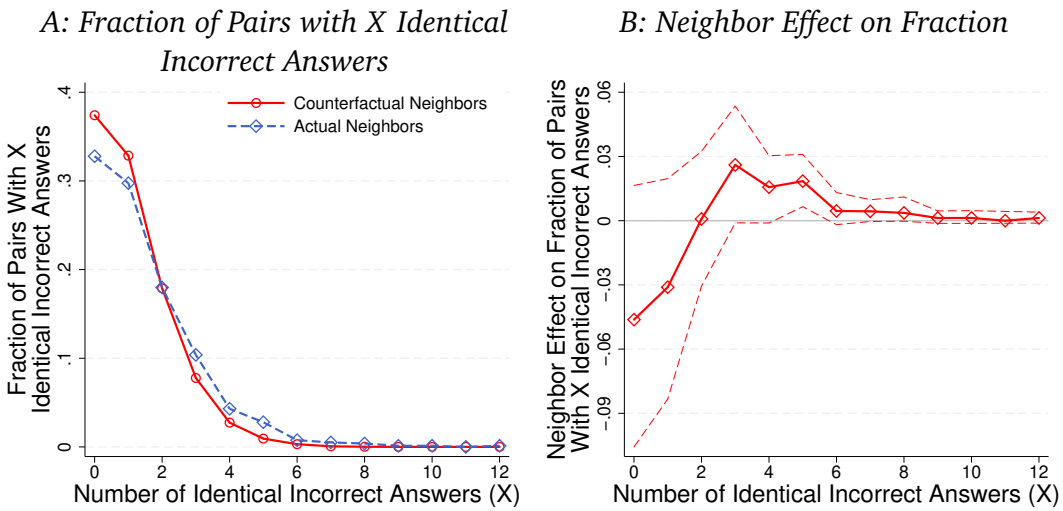
Notes: This figure examines how the students' ability (proxied by high-school performance) relates to their cheating behavior (under baseline monitoring). To construct this figure, we use model (2) to estimate the effect of being a pair of actual neighbors on the probability that two students give identical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

Figure B12: Grade Heterogeneity under Close Monitoring



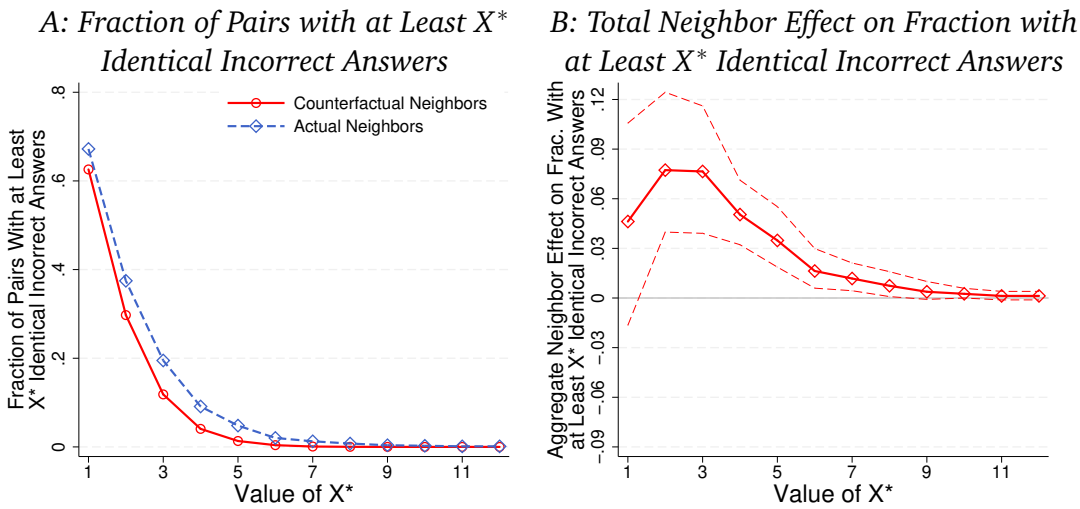
Notes: This figure examines how the students' ability (proxied by high-school performance) relates to their cheating behavior (under close monitoring). To construct this figure, we use model (2) to estimate the effect of being a pair of actual neighbors on the probability that two students give identical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

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



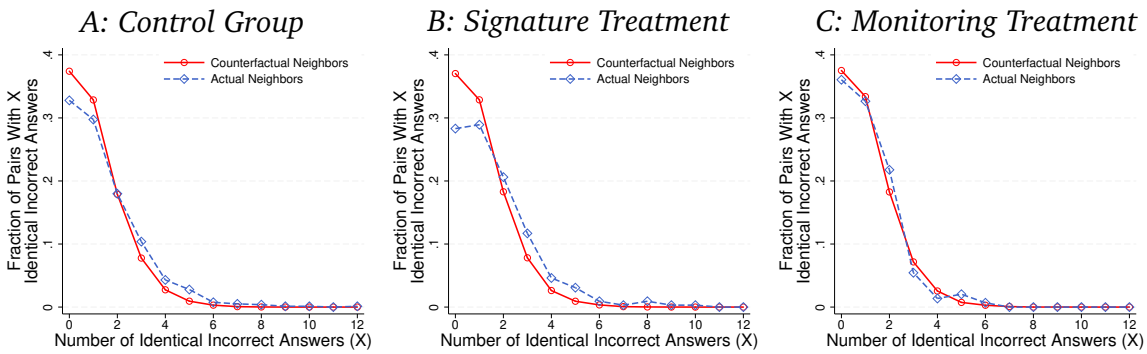
Notes: This figure shows how cheating shifts the distribution of identical incorrect answers for $X = 0$ to $X = 12$. Panel *A* depicts the distribution of identical incorrect answers for counterfactual neighbor pairs (solid red line) and actual neighbor pairs (dashed blue line). Panel *B* shows the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

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



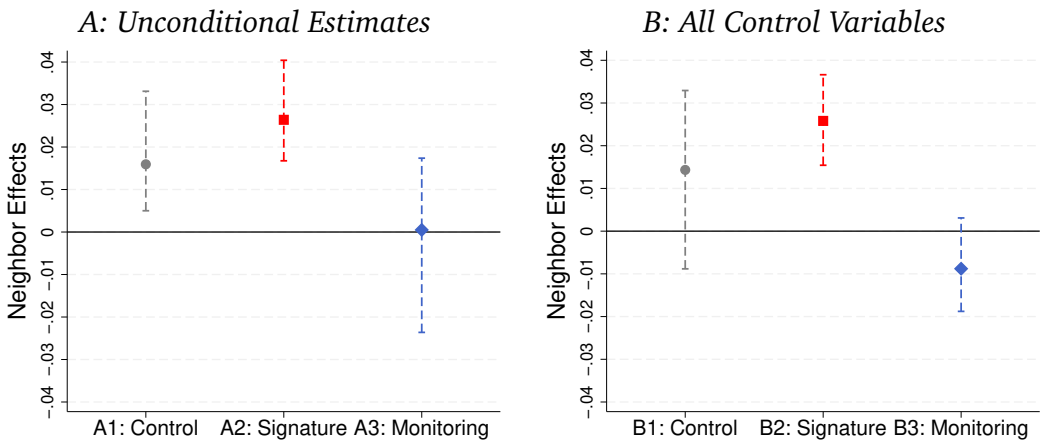
Notes: This figure shows how cheating shifts the fraction of pairs with more than X^* identical incorrect answers. The solid red line in Panel *A* depicts the fraction for counterfactual neighbor pairs as a function of X^*: $\sum_{X=X^*}^{30} \widetilde{f}^X$. The dashed blue line shows the corresponding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

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



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

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



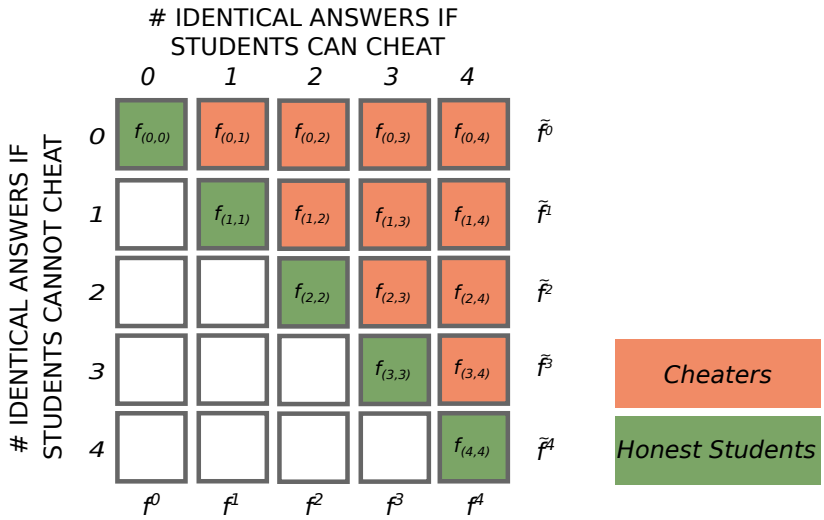
Notes: This figure shows the average neighbor effect on identical (correct and incorrect) answers for the control group (gray circles), the signature treatment (red squares), and the monitoring treatment (blue diamonds). To construct this figure, we estimate model (3) and, based on this model, predict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

C Lower Bound for the Share of Cheating Pairs

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

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

Figure C1: Transition Matrix



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

actual and counterfactual neighbors differs.[1] By definition, plagiarism increases the similarity in the students' answers. Consequently, there are no pairs that share more incorrect answers when they cannot than when they can cheat.

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

Observable Entities. Empirically, we do not directly observe all the elements of the transition matrix, which complicates the estimation for the share of cheaters. For example, we do not know how many pairs would increase their identical incorrect answers from zero to one when they can cheat. However, we can observe or estimate several entities that allow us to estimate a lower bound for the share of cheaters. First, we directly observe f^X, the fraction of actual neighbors who share X incorrect answers when they can cheat. In the transition matrix, f^X is the column sum for column X.[2] Second, using the empirical approach described in Section III.B, we are also able to estimate $\tilde{f}^X$, the fraction of pairs sharing X incorrect answers in the scenario without cheating. This estimate approximates the row sum for row X. Third, given that we know f^X and $\tilde{f}^X$, we can estimate the neighbor effect on the distribution at value X:

$$NED^X = f^X - \tilde{f}^X. \tag{1}$$

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

$$TNE^{X^*} = \sum_{X=X^*} (f^X - \tilde{f}^X). \tag{2}$$

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

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

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

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

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

By plugging the definitions for f^1 to f^4 and $\tilde{f}^1$ to $\tilde{f}^4$ into equation (3)[3] and by rearranging terms, we can derive the expression:

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

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

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

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

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

Estimating a Lower Bound. Building on the previously discussed insights, our procedure to estimate a lower bound for the share of cheaters proceeds in two steps. First, we estimate the set of total aggregated neighbor effects $TNE = \{TNE^1, TNE^2, ..., TNE^{29}\}$. The estimates rely on our identifying assumption, stating that plagiarism is the only systematic reason why the similarity in the answers of actual neighbors differs from that in the answers of counterfactual neighbors.[4] Second, our lower bound for the share of cheaters corresponds the total aggregated neighbor effect TNE^{X^*} that maximizes TNE.

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

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

Limitation. The main limitation of our distribution-based analysis is that we are only able to estimate lower bounds. The reason is that the distributions do not reveal each

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

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

pair's position in the transition matrix. Thus, based on distributions, we can neither identify the elements of the transition matrix nor the share of cheaters. Instead, as discussed, the total aggregated neighbor effects only identify subsets of cheaters. Appendix D addresses this limitation. It proposes a method that allows us to identify cheating at the pair level.

D Share of Cheaters: Alternative Test

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

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

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

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

Results. We calculate the share of cheaters for multiple specifications. First, we bound the false-discovery rate to 5% and apply the testing procedure to identical (correct and incorrect) answers. Second, we restrict the rate to 1% and consider the same outcome. For both of these two specifications, we identify 8.0% of all pairs as cheaters. Third, we consider identical incorrect answers instead of all identical answers and, again, apply both false-discovery rates. In this case, we find a slight increase in the share of cheaters to 9.2%. This result is in line with our previous observation that identical incorrect answers contain more precise traces of cheating. We conclude that the results are fairly in line with our lower-bound estimate.

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

E Suggestive Evidence on Channels

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

E.A Experimental Design

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

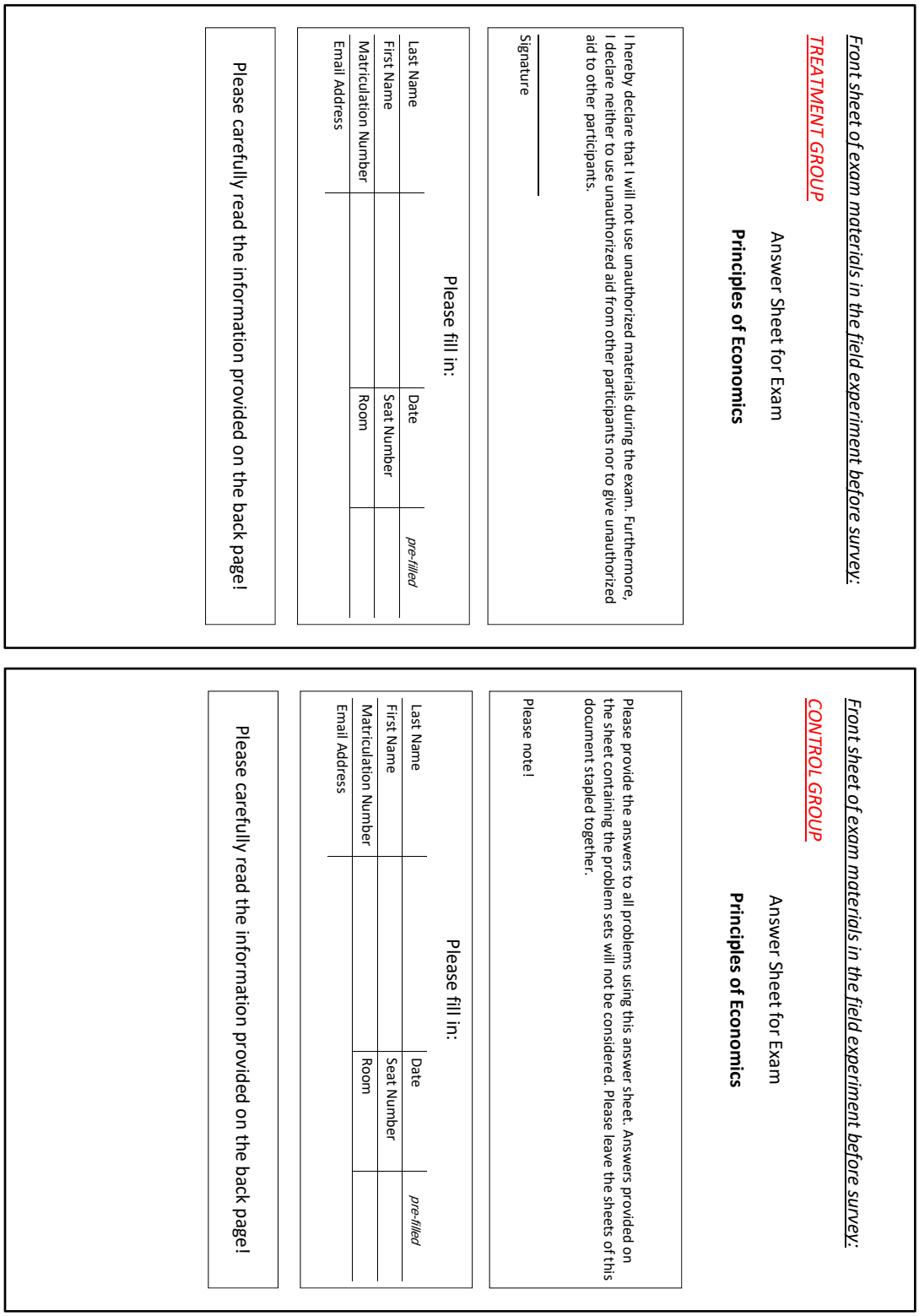
Signature Treatment and Control Group: Details. We implemented our follow-up experiment in the first exam (principles of economics). As in the initial experiment, the follow-up experiment evenly split the examinees into a control group and a signature treatment. Furthermore, the signature treatment implemented an honesty declaration that was identical to the one in the initial experiment. However, in contrast to the initial experiment, we randomly assigned the treatment status within the lecture halls. Therefore, not all the students in one hall did receive the same treatment. Instead, some students were in the control group and others in the signature treatment.[6] The paragraph "Why Follow-Up Experiments?" motivates and discusses this design element in more detail.

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

Survey. Two hours after the exams, we invited the examinees to participate in an online survey (via the mailing list). Students who accepted the invitation were redirected to the

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

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



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

TREATMENT GROUP

Answer Sheet for Exam

Principles of Economics

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

Signature

Please fill in:

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

Please carefully read the information provided on the back page!

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

CONTROL GROUP

Answer Sheet for Exam

Principles of Economics

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

Please note!

Please fill in:

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

Please carefully read the information provided on the back page!

welcome page of the online survey. This page informed participants that the survey's goal was to measure "how students generally perceive exams at the university." It also asked

participants to think about their last exam when answering the questions.[7] To prevent that students foresaw our goal to study the impact of the honesty declaration on their survey responses, we did not refer to the previous exam on principles of economics at any point during the survey. Furthermore, answering the survey took about five minutes, and participants received a flat payoff of €3.50.

Table E1: Post-Exam Survey: Questions

Perceived Sanction
S1 & S2: Imagine the supervising staff in your last exam had witnessed how one participant *copies answers from other participants [uses unauthorized materials (like, for instance, a smartphone)]*. What do you think would be the likely consequence for this student?
Perceived Detection Probability
D1 & D2: Think back to your last exam, and imagine 100 participants who *try to copy at least one answer from other participants [use unauthorized materials (like, for instance, a smartphone) to answer at least one question]*. What do you think, how many of those 100 students would have been caught? Please state a number between 0 and 100.
Descriptive Norm
N1 & N2: Think back to your last exam, and imagine a group of 100 participants. What do you think, how many of those have *copied at least one answer from other participants [used unauthorized materials (like, for instance, a smartphone) to answer at least one question]*? Please state a number between 0 and 100.
N3 & N4: Think back to your last exam, and imagine the supervising staff had left the exam hall for a few minutes. What do you think, how many of 100 participants would have *copied at least one answer from other participants in the meanwhile [used unauthorized materials (like, for instance, a smartphone) to answer at least one question in the meanwhile]*? Please state a number between 0 and 100.

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

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

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

Table E2: Post-Exam Survey: Balancing Checks

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

Notes: This table shows balancing checks. It reports mean values of the covariates separately for the control group (Column (1)) and the signature treatment (Column (2)). Moreover, Column (3) shows the difference in the mean values between the signature treatment and the control group with standard errors in parentheses. Panel *A* focuses on the sample of students participating in the repetition of the field experiment. Panel *B* considers the sample of individuals who participated in post-exam survey. Math proficiency is only available for 775 out of the 1060 students (Panel A) and 83 out of 103 survey participants (Panel B).

Sample: Details. We focused on two cohorts of students who took the exam on principles of economics in two years after our initial experiment.[8] In total, the two cohorts consisted of 1060 students of which 233 signed up for the mailing list before the exam. Ultimately, 103 students completed the survey. Panel *A* in Table E2 shows that the signature treatment and the control group are balanced in observable characteristics. Moreover, the 103 survey participants have similar observable characteristics as non-participants (see Table E3). The only significant sample imbalance is that participants have high-school GPAs which are 0.19 grade points (or 0.31 standard deviations) better than non-participants.[9] Furthermore, in the sample of survey participants, the signature treatment

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

[9]To probe the robustness of the survey evidence regarding this type of sample imbalance, we also estimated models that reweigh the individual observations such that the first moments of all characteristics are

($N = 48$) and the control group ($N = 55$) are well-balanced in all the observable characteristics (see Panel B in Table E2).

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

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

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

Why Follow-Up Experiments? One may wonder why we needed to implement follow-up experiments to study channels. The reason is that two complications forbid us to analyze channels in our initial experiment. First, due to local exam regulations, we were not allowed to ask survey questions during the exam. After we implemented our initial experiment, the department of economics, however, established the aforementioned mailing list for academic surveys. This newly established survey allowed us to invite students who took part in the follow-up experiment to our online survey. Second, the fact that we had to survey students after the exam complicated identification. This holds especially in our initial experiment that randomized the treatments across lecture halls. To understand the potential issue of such a design, consider, for example, perceived norms as a channel. If the signature treatment increased cheating, post-exam questions on norms may reflect that students in the signature-treatment halls observe more cheating than students in control-group halls, instead of reflecting shifts in the perceived norm. This would lead us to overestimate the effect of the request on perceived norms. To tackle this issue, we adjusted the sampling scheme and, as previously mentioned, randomly assigned the signature treatment within lecture halls. This design element ensures that, on average, students in both groups experienced the same level of cheating by peers.

Drawbacks. Our design has two drawbacks. First, one obvious limitation is the limited sample size. Not all students signed up for the mailing list, and survey participation conditional on being registered is voluntary. Second, our design forbids us to reestimate the effect of the request to sign the honesty declaration on cheating. The reason is that

identical to the population estimates. The results are virtually unchanged to those reported subsequently.

Table E4: Post-Exam Survey: Perceived Sanctions

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

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

the treatment status differs between neighbors. Similarities in their answers, thus, reflect a mix of cheating in both conditions.

E.B Results

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

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

$$Y_{ih} = \gamma_0 + \gamma_S S_{ih} + X_{ih}\gamma_X + \pi_h + u_{ih}, \tag{4}$$

where Y_{ih} is the stated perception of student i seated in hall h, S_{ih} is an indicator for the signature treatment, X_{ih} is a vector of student controls (age, gender, and high-school GPA), and π_h absorbs exam-hall fixed effects. Regarding inference, we provide heteroscedasticity-consistent and wild-cluster-bootstrap p-values (17 clusters at the exam hall-level). In addition to the treatment effects for the individual outcomes, we also report average standardized effects according to Kling *et al.* (2004) and Clingingsmith *et al.* (2009) and exploit Mann-Whitney-U-Tests to non-parametrically test for treatment differences.

Columns (1) to (3) in Table E5 present the main results. The columns show that the signature treatment neither shifts the perceived detection probability in case of copying

Table E5: Post-Exam Survey: Treatment Effects

	Detection Probability Copying	Detection Probability Unauthorized Materials	Average Stand. Effect (1) & (2)	% Other Students Copying	% Others Using Unauthorized Materials	Average Stand. Effect (4) & (5)	Social Norm Copying	Social Norm Unauthorized Materials	Average Stand. Effect (7) & (8)	Average Stand. Effect (4),(5),(7),(8)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Effect of Signature Treatment	-6.0	-2.5	-0.13	4.5	3.6	0.61	14.9	19.9	0.54	0.58
p-value, robust	[0.400]	[0.727]	[0.489]	[0.228]	[0.010]***	[0.012]**	[0.026]**	[0.004]***	[0.001]***	[0.001]***
p-value, hall cluster, wild bootstrap	[0.368]	[0.710]		[0.326]	[0.024]**		[0.045]**	[0.021]**		
p-value, Mann-Whitney-U-Test	[0.686]	[0.829]		[0.185]	[0.041]**		[0.053]*	[0.029]**		
Control Group Mean	31.6	31.4		8.6	3.4		52.2	43.4		
Number of Observations	103	103		103	103		103	103		

Notes: The table reports the effects of the signature treatment on students' responses in the post-exam survey. The estimates are derived from OLS regressions using gender, age, high-school GPA, and exam-hall fixed effects as additional controls. We report the following types of *p*-values [in brackets]: (a) heteroscedasticity robust *p*-values, (b) hall-cluster-robust *p*-values based on a wild-cluster-bootstrap procedure that accounts for the small number of cluster (Cameron *et al.*, 2008), (c) *p*-values for Mann-Whitney-U-Tests, and (d) robust *p*-values for average standardized effects following Kling *et al.* (2004) and Clingingsmith *et al.* (2009). Dependent variables in Column (1): perceived detection probability (in percent) if copying from a neighbor. Column (2): perceived detection probability (in percent) if using unauthorized materials (like smartphone, etc). Column (4): perceived share (in percent) of students copying at least one answer. Column (5): perceived share (in percent) of students using unauthorized materials. Column (7): perceived share of students (in percent) that would copy at least one answer in case of no supervision. Column (8): perceived share of students (in percent) that would use unauthorized materials in case of no supervision. See Table E1 for the exact wording of the survey questions.

nor the one in case of using unauthorized materials. Taken together with the fact that the average student is well informed about the actual sanction, the absence of a significant treatment effect suggests that the signature treatment did also not significantly shift the participants' expected sanction for cheating.

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

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

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

References

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

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

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

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