A Field Experiment on Intertemporal Enforcement Spillovers

Tobias Cagala, Ulrich Glogowsky, Johannes Rincke*

July 9, 2014

Abstract

Enforcement can affect compliance directly and indirectly, through spillovers. We study intertemporal enforcement spillovers by conducting a randomized field experiment in a university exam. The initial phase of the experiment is the exam itself. We induce variation in enforcement by randomly assigning students to rooms with different monitoring levels. The second phase is post-exam. In this phase, all students are subject to the same low level of monitoring. Our outcome variable is the probability that students steal a pen in the post-exam phase. We find that enforcement in the exam phase has a strong intertemporal spillover on compliance in the post-exam phase: students subject to a high monitoring level in the initial phase are about 33 percent less likely to steal the pen than students subject to low initial monitoring.

JEL Codes: C93, D03, D83

Keywords: enforcement, spillover, compliance

*Cagala: University of Erlangen-Nuremberg (tobias.cagala@fau.de); Glogowsky: University of Erlangen-Nuremberg (ulrich.glogowsky@fau.de); Rincke: University of Erlangen-Nuremberg (johannes.rincke@fau.de). We thank Tim Friehe, Christian Traxler and numerous seminar participants for helpful comments. We are grateful for financial support from the Emerging Field Initiative at University of Erlangen-Nuremberg. All errors are our own.

1 Introduction

Since Becker (1968), the literature on individual compliance with rules and norms has highlighted the importance of direct effects of enforcement. It is well established that across various contexts including criminal activity and tax evasion, enforcement is an important driver of compliance (Levitt, 1997, 2002; Di Tella and Schargrodsky, 2004; Slemrod *et al.*, 2001; Kleven *et al.*, 2011; Fellner *et al.*, 2013; Dwenger *et al.*, 2014).[1] The exact mechanisms through which enforcement actually works are, however, often unclear. This holds in particular for indirect effects like cross-sectional or intertemporal spillovers. Among the few available studies, Rincke and Traxler (2011) and Pomeranz (2013) use field data to identify between-subject spillovers, while Bruttel and Friehe (2014) consider intertemporal spillovers in the lab.

This study provides novel evidence on intertemporal (within-subject) enforcement spillovers from a randomized field experiment. It was implemented in a university exam and had two phases. In the exam phase, students were randomly assigned to two treatments which differed in the intensity of monitoring during the exam. In both treatments, students were provided with a high-quality pen. In the post-exam phase, exam materials were re-collected, and students left the room. In this phase, all students were subject to the same low level of monitoring. The outcome studied is the probability that students steal the pen in the post-exam phase. We find that enforcement in the exam phase had a strong intertemporal spillover: students subject to the high monitoring level in the initial phase were much less likely to steal the pen than students subject to low initial monitoring. This is consistent with perceptions of initial enforcement serving as an anchor when subjects assess the risk associated with non-compliance at subsequent stages (Tversky and Kahneman, 1974).

We describe our design and data in Section 2. Results are presented in Section 3. Section 4 concludes.

2 Experimental Design and Data

2.1 Setting

The experiment was conducted in February 2013 at the department for economics and business administration of a German university.[2] The sample consists of 766 freshmen who were randomly assigned into treatments using strata defined by gender and grade of university entrance qualification. We also randomized the allocation of students to seats within rooms.

All conditions other than the level of monitoring were highly standardized across treatments. Supervising staff in all rooms behaved according to a written exam schedule, including the exact wording for instructions to be read out to students.[3] The issue of no-shows (leading to

[1] A recent literature also addresses the crowding out of compliance by enforcement (Boyer *et al.*, 2014; Gangl *et al.*, 2014).

[2] All interventions were agreed upon by the department's examination board.

[3] In two out of eight rooms, in addition to standard oral instructions given in all rooms that students caught

unintended differences in the number of supervisors per student) was solved by overbooking rooms. Overbooking enabled us to draw from a room-specific pool of students to fill seats that otherwise would have remained empty due to no-shows.[4] To standardize effective monitoring during the exam, supervising staff was instructed to stay at predetermined locations. Effective monitoring was thus easily observable to students.

High-quality pens with a retail price of about €3 were distributed to students' pre-assigned seats before the exam. Supervisors (truthfully) announced that using the pen would facilitate the automated scanner-based evaluation of multiple-choice answer sheets.[5] Instructions read out at the beginning and after the end of the exam informed students to leave the pen at their seat for later recollection. This ensures that students did not perceive the pen as a gift.

2.2 Exam Phase and Implementation of Treatments

The experiment involved an exam phase and a post-exam phase. The timing in the exam phase was as follows:

1. Students enter room and take their seat.
2. Supervisors read out instructions.
3. Supervisors take pre-assigned positions and students start working.

We implemented two treatments to induce variation in the level of monitoring during the exam. In the *Low Monitoring Treatment* ($N = 541$), we assigned the minimum number of supervisors to rooms according to local exam regulations (1.11 supervisors per 50 students on average). In the *High Monitoring Treatment* ($N = 225$), monitoring was much stricter (6.00 supervisors per 50 students).

2.3 Post-Exam Phase

In the post-exam phase, the timing was as follows:

1. Supervisors announce the end of the examination.
2. The monitoring level in High Monitoring rooms is reduced to the level maintained in Low Monitoring rooms. In each of the High Monitoring rooms, a corresponding number of supervisors leave the room.
3. Supervisors collect exam materials.
4. Students leave the room.

cheating will fail the exam, students were asked to sign a no-cheating declaration. We carefully checked all available outcomes for a potential effect of the declaration, including tests for traces of cheating like above-normal similarity in performance between direct neighbors. We did not find any difference between students who were asked to sign the declaration and those who did not. See Cagala *et al.* (2014) for details.

[4]Apart from our sample, 235 additional students took part in the exam. Students originally assigned to a given room who could not be seated due to overbooking were relegated to other rooms that were designated to accommodate these students. Subjects seated in these rooms are not part of our sample.

[5]We checked all answer sheets and found that 96% of students apparently used the provided pen to check the multiple-choice boxes.

We took care as to maximize the saliency of the change in monitoring conditions in the High Monitoring Treatment. When supervisors announced that the exam was over, students remained at their seats and awaited further instructions. It was at this point in time (i.e., with no other activity going on) that the additional supervisors in the High Monitoring Treatment left the room. The remaining supervisors then collected the exam materials. In all rooms, they did so under the instruction to focus exclusively on exam materials and not to respond to anything related to the pens.

2.4 Data

We link administrative data on individual student characteristics (age, gender, field of study, and grade of university admission qualification) with data on the location of pens after students had left the examination rooms.[6] Standard randomization checks demonstrate that our treatment groups are well balanced across all observable student characteristics.

3 Evidence on Intertemporal Enforcement Spillovers

We report estimates for the impact of exam phase monitoring on the probability for students stealing a pen in the post-exam phase, when monitoring was identical across both treatments. First, we estimate the spillover using only the *aggregate* share of missing pens by treatment group. Our baseline measure for the share of missing pens comes from the Low Monitoring Treatment, where 75 out of the total number of 541 pens were missing. This gives a baseline measure of non-compliance under low monitoring of 13.9 percent. In the High Monitoring Treatment, only 21 out of a total of 225 pens were missing (9.3 percent). Our most conservative estimate of the intertemporal enforcement spillover is thus a 32.7 percent drop in non-compliance relative to the low monitoring baseline. A nonparametric one-tailed Fisher's exact test confirms the existence of a sizable intertemporal spillover ($p = 0.052$).

Our second step is to estimate the spillover using individual data. This requires the assumption that any measurement error in the outcome is unrelated to treatment assignment. Given the careful randomization, we are confident that this assumption holds.[7] Table 1 summarizes the evidence from a total of six individual-level OLS regressions.

[6] In a few cases, pens were not left *at* seats, but left *between* two seats. In these cases, we randomly assign the pen to one of the two seats.

[7] Measurement error in our outcome variable might arise from two types of behavior. First, students might take the pen they were provided with but then leave it at a different seat before leaving the room. This behavior would imply we should find *more* than one pen at some seats. This was, however, not the case. Second, students could leave the pen at their seat but take a pen from another seat. We address this issue later on when discussing the issue of spatial correlations.

Table 1: Estimates of Intertemporal Enforcement Spillover

	Spillover Treatment Effect (%)		Spillover Effect of Monitoring Intensity	
	Effect	p-Value	Effect	p-Value
	(1)	(2)	(3)	(4)
A. No controls	-32.7	0.064	-0.925	0.066
	(17.6)		(0.503)	
B. Controlling for Strata Variables	-33.1	0.061	-0.938	0.063
(University Admission Grade & Gender)	(17.6)		(0.503)	
C. Controlling for Strata Variables,	-35.3	0.046	-1.00	0.047
Age & Field of Study	(17.7)		(0.504)	
Non-Compliance in Omitted Reference Group	13.9%		-	
Number of observations	766		766	

Notes: The table reports coefficients, heteroscedasticity-robust standard errors (in parentheses), and corresponding p-Values from OLS regressions. Column (1) shows the results for the spillover treatment effect in percent (treatment effect on probability for missing pen relative to share of missing pens in omitted reference group). Column (3) reports estimations with the monitoring intensity (number of supervisors per 50 students) in the exam phase as the main explanatory variable. In column (3), the coefficient gives the effect of an additional supervisor per 50 students on the probability for a missing pen in percentage points. Panel A shows results for estimations without any controls. Panel B reports the same estimations while controlling for strata variables (dummies for university admission grade (four bins) and gender). In Panel C, we control for strata variables plus age and field of study. The bottom of the table reports the number of observations and the baseline non-compliance rate (percentage of missing pens) in the Low Monitoring group.

Column (1) shows results for the spillover treatment effect in percent. This is derived by dividing the treatment effect on the probability for a missing pen by the share of missing pens in the Low Monitoring Treatment group. Panel A shows this relative treatment effect without further controls, yielding the baseline estimate of a 32.7 percent drop in non-compliance (probability for stealing pen) relative to the Low Monitoring Treatment group. Once we include individual controls, the point estimates get slightly larger (in absolute terms): Panel B reports the treatment effect when controlling for strata variables, and Panel C for using strata variables plus *age* and *field of study* as additional controls. Column (1), Panel C suggests that the High Monitoring Treatment has triggered a 35.3 percent drop in the probability for stealing a pen relative to the Low Monitoring Treatment. Our estimates thus identify an economically and statistically significant intertemporal enforcement spillover.

Column (3) reports estimations with the monitoring intensity in the exam phase, measured as number of supervisors per 50 students, being the main explanatory variable. Given random assignment into treatments, the monitoring intensity is an accurate continuous measure of exogenous experimental variation in actual enforcement. Across panels A to C, the pattern of estimation results is very similar to that in column (1). With all controls included, we find that

an additional supervisor per 50 students in the exam phase lowers the probability of a missing pen by one percentage point.

We probed our results by a series of robustness checks. First, spatial correlations in our outcome would violate the independence assumption. Such correlation could result from students taking several pens from seats next to their own, or from cross-sectional behavioral spillovers, i.e. from students adjusting their behavior after observing other students taking the pen, possibly due to conformity motives (Bernheim, 1994). Room-specific permutation tests for spatial correlation, however, do not reject the null of no spatial correlation for seven out of eight treatment rooms. Even when repeating all regressions while excluding observations from the room where the null was rejected, we found all our previous results confirmed (see supplemental materials for details). Second, we checked that using randomization inference which is robust to correlations within clusters (i.e., rooms) does not affect our findings. Third, we checked that students who were assigned to seats that otherwise would have remained empty due to no-shows did not behave differently from students with pre-assigned seat numbers. Fourth, it is conceivable that feeling more or less self-confident immediately after the exam translates into differences in cheating behavior. However, when controlling for exam performance (percent correctly solved problems), we found the performance indicator to be insignificant, and all our findings unchanged. Finally, we obtained very similar estimation results when fitting probit models instead of linear probability models.

4 Conclusion

Intertemporal enforcement spillovers can work through different channels, and it is difficult to distinguish them empirically. Bruttel and Friehe (2014) present related lab evidence and conclude that loss aversion drives individual compliance decisions. Alternatively, intertemporal enforcement spillovers are also consistent with anchoring (Tversky and Kahneman, 1974): the initial level of enforcement experienced by an individual could serve as an anchor, biasing individuals' subjective estimates of actual enforcement in later periods. In our experiment, anchoring would lead students in the High Monitoring Treatment to under-adjust their estimate of the actual monitoring intensity once the 'excess monitoring' relative to the baseline is removed. A standard response to the perceived level of monitoring would then explain the spillover.

The intertemporal spillover identified in this study is of a short-term nature. It remains for future research to study intertemporal spillovers in contexts where differences between short-term and long-term effects could be relevant. One possible implication of spillovers fading out over time could be that short-term evaluations of policy shifts towards increased deterrence in the context of crime, tax evasion, etc. would provide lower bound estimates of actual long-run effects.

References

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

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

Boyer, P., Dwenger, N. and Rincke, J. (2014). Do Taxes Crowd Out Intrinsic Motivation?, unpublished.

Bruttel, L. and Friehe, T. (2014). On the Path Dependence of Tax Compliance. *European Economic Review*, **65**, 90–107.

Cagala, T., Glogowsky, U. and Rincke, J. (2014). Does Commitment to Rules Increase Compliance? Combined Laboratory and Field-Experimental Evidence, unpublished.

Di Tella, R. and Schargrodsky, E. (2004). Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack. *American Economic Review*, **94**, 115–133.

Dwenger, N., Kleven, H. J., Rasul, I. and Rincke, J. (2014). Extrinsic and Intrinsic Motivations for Tax Compliance: Evidence from a Field Experiment in Germany, unpublished.

Fellner, G., Sausgruber, R. and Traxler, C. (2013). Testing Enforcement Strategies in the Field: Threat, Moral appeal and Social Information. *Journal of the European Economic Association*, **11** (3), 634–660.

Gangl, K., Torgler, B., Kirchler, E. and Hofmann, E. (2014). Effects of Supervision on Tax Compliance: Evidence from a Field Experiment in Austria. *Economics Letters*, **123**, 378–382.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S. and Saez, E. (2011). Unwilling or Unable to Cheat? Evidence From a Tax Audit Experiment in Denmark. *Econometrica*, **79**, 651–692.

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

— (2002). Using Electoral Cycles in Police Hiring to Estimate the Effects of Police on Crime: Reply. *American Economic Review*, **92**, 1244–1250.

Pomeranz, D. (2013). *No Taxation without Information: Deterrence and Self-Enforcement in the Value Added Tax*. NBER Working Papers 19199, National Bureau of Economic Research, Inc.

Rincke, J. and Traxler, C. (2011). Enforcement Spillovers. *Review of Economics and Statistics*, **93**, 1224–1234.

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

Tversky, A. and Kahneman, D. (1974). Judgment Under Uncertainty: Heuristics and Biases. *Science*, **185**, 1124–1131.

Supplemental Material (For Online Publication Only)

Descriptive Statistics

Table 2: Balancing Checks

	T1 Low Monitoring Mean (1)	T2 High Monitoring Mean (2)	Difference Mean (3)	Difference p-Value (3)
Gender (Female = 1)	0.545	0.502	0.043	0.278
			(0.040)	
University Admission Grade	2.47	2.50	-0.025	0.568
			(0.044)	
Age	19.6	19.6	0.014	0.881
			(0.096)	
Field of Study (Economics & Sociology = 1)	0.065	0.089	-0.024	0.266
			(0.022)	

Notes: This table shows means of individual student characteristics and balancing checks. Columns 1 and 2 show treatment-specific means. Column 3 shows the difference in means with heteroscedasticity-robust standard errors in parentheses. The University Admission Grade is the overall grade of the university admission qualification, typically obtained in High School, ranging from 1.0 (outstanding) to 4.0 (pass). Age is derived assuming that students achieve university admission at the age of 19. 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*. The sample consists of the 766 students in the experiment. *University Admission Grade* and *Gender* were used for stratification.

Table 3: Monitoring and Compliance Across Examination Rooms

Room	Treatment	Number of Supervisors	Number of Students	Monitoring Intensity	Share of Missing Pens (%)
(1)	(2)	(3)	(4)	(5)	(6)
1	T1	4	199	1.01	15.6
2	T1	2	113	0.88	13.3
3	T1	2	95	1.05	14.7
4	T1	2	76	1.32	11.8
5	T1	2	58	1.72	10.3
	Mean Low Monitoring Treatment:			1.11	13.9
6	T2	15	119	6.30	10.1
7	T2	6	55	5.45	5.5
8	T2	6	51	5.88	11.8
	Mean High Monitoring Treatment:			6.00	9.33

Notes: The table summarizes the monitoring regimes and compliance for all eight rooms in the experiment. The monitoring intensity is measured as number of supervisors per 50 students. In the smallest Low Monitoring rooms (rooms 4 and 5), local exam regulations requiring a minimum of two supervisors overseeing each examination room prevented us from implementing even lower monitoring intensities.

Robustness Against Spatial Correlations

We ran a series of room-specific permutation tests for spatial correlation in our outcome variable (indicator for missing peen). The test is based on an index capturing the spatial similarity in outcomes between direct neighbors.[8] To obtain the test statistic, we derive the resample distribution of the index under the null of no spatial correlation. This is done by 10,000 repetitions of a procedure that, in each run, randomly reassigns seats within rooms (without replacement) and calculates the index for the reshuffled seating plan.

The permutation tests do not reject the null of no spatial correlation in the probability for missing pens for seven out of eight rooms. The only room where the test rejects the null is room no. 6, where we implemented the High Monitoring Treatment. In a second step of our robustness check, we repeated all regressions reported in the paper (Table 1) while excluding observations from this room. Table 4 reports the results. Even though excluding room no. 6 excludes 119 observations (and therefore leaves us with less power), we find all our previous

[8]Formally, the index is derived as

$$I = \frac{\sum_i \sum_{j \neq i} w_{ij} 1[P_i = P_j]}{\sum_i \sum_{j \neq i} w_{ij}}, \quad i, j \in \{1, \ldots, N\}. \tag{1}$$

The spatial weight w_{ij} takes the value 1 if students i and j are direct (row- or column-wise) neighbors and 0 otherwise. $1[\cdot]$ is the indicator function taking the value 1 if the outcome P (pen present or missing) between i and j is equal, and 0 otherwise. Intuitively, the index measures the probability for identical outcomes between neighbors.

results confirmed. In fact, the point estimates are higher compared to the estimations using the full sample. At the same time, statistical significance is only marginally affected.

Table 4: Estimates of Intertemporal Enforcement Spillover Excluding Room No. 6

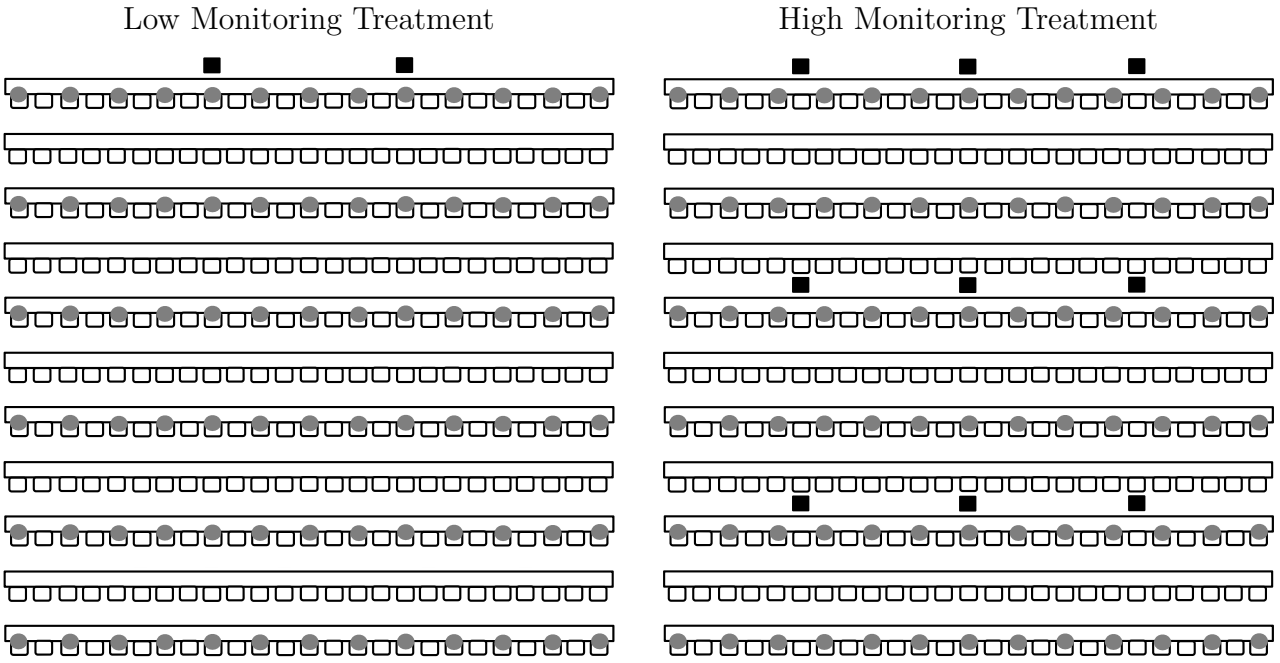
	Spillover Treatment Effect (%)		Spillover Effect of Monitoring Intensity	
	Effect	p-Value	Effect	p-Value
	(1)	(2)	(3)	(4)
A. No controls	-38.8	0.083	-1.22	0.078
	(22.3)		(0.688)	
B. Controlling for Strata Variables	-39.7	0.074	-1.24	0.069
(University Admission Grade & Gender)	(22.1)		(0.683)	
C. Controlling for Strata Variables,	-42.1	0.059	-1.33	0.054
Age & Field of Study	(22.3)		(0.687)	
Non-Compliance in Omitted Reference Group	13.9%		-	
Number of observations	647		647	

Notes: The table reports coefficients, heteroscedasticity-robust standard errors (in parentheses), and corresponding p-values from OLS regressions. The difference to Table 1 in the paper is that we exclude observations from room no. 6, as a spatial permutation test rejects the null of no spatial correlation for this room. Column (1) shows results for the spillover treatment effect in percent (treatment effect on probability for missing pen relative to share of missing pens in omitted reference group), whereas column (3) reports results with the monitoring intensity (number of supervisors per 50 students) in the exam phase being the main explanatory variable. In column (3), the coefficient gives the effect of an additional supervisor per 50 students on the probability for a missing pen in percentage points. Panel A shows results for estimations without any controls. Panel B reports the same estimations while controlling for strata variables (defined by university admission grade (four bins) and gender). In Panel C, we control for strata variables plus age and field of study. The bottom of the table reports the number of observations and the baseline non-compliance rate (percentage of missing pens) in the Low Monitoring group.

Supplemental Figures on Implementation Issues

Figure 1: The High-Quality Pen

Figure 2: Enforcement During the Exam Period



Notes: This figure gives a stylized illustration of monitoring during the exam phase in the *Low Monitoring Treatment* (left panel) and the *High Monitoring Treatment* (right panel). Gray dots represent students; black squares represent supervisors.