Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Spillover Effects of Monitoring: A Field Experiment

Michèle Belot, Marina Schröder

To cite this article:
Michèle Belot, Marina Schröder (2016) The Spillover Effects of Monitoring: A Field Experiment. Management Science 62(1):37-45. http://dx.doi.org/10.1287/mnsc.2014.2089

MANAGEMENT SCIENCE
Vol. 62, No. 1, January 2016, pp. 37–45
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

informs®

http://dx.doi.org/10.1287/mnsc.2014.2089
© 2016 INFORMS

The Spillover Effects of Monitoring: A Field Experiment

Michèle Belot
School of Economics, University of Edinburgh, Edinburgh EH8 9JT, United Kingdom, michele.belot@ed.ac.uk

Marina Schröder
Faculty of Management, Economics and Social Sciences, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany, marina.schroeder@uni-koeln.de

We provide field experimental evidence of the effects of monitoring in a context where productivity is multidimensional and only one dimension is monitored and incentivized. We hire students to do a job for us. The job consists of identifying euro coins. We study the direct effects of monitoring and penalizing mistakes on work quality and evaluate spillovers on unmonitored dimensions of productivity (punctuality and theft). We find that monitoring improves work quality only if incentives are harsh, but substantially reduces punctuality irrespectively of the associated incentives. Monitoring does not affect theft, with 10% of participants stealing overall. Our findings are supportive of a reciprocity mechanism, whereby workers retaliate for being distrusted.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mnsc.2014.2089.

Keywords: counterproductive behavior; monitoring; field experiment
History: Received November 5, 2013; accepted August 23, 2014, by Uri Gneezy, behavioral economics. Published online in *Articles in Advance* March 13, 2015.

1. Introduction

Experts estimate that, globally, occupational fraud causes annual losses of more than $3.5 trillion (Association of Certified Fraud Examiners 2012). The question is what an organization can do to prevent such behavior. One straightforward instrument regularly applied in practice is to monitor workers and punish them if they do not comply (or reward them if they do). But are such measures effective? There is experimental evidence that monitoring and incentivizing may actually backfire (for reviews of this literature, see Frey 1993, Falk and Kosfeld 2006, Frey and Jegen 2001). However, the evidence is so far limited to situations where productivity is unidimensional, such as the number of units produced or sold, performance at a test, or monetary transfers in an experimental game (see, e.g., Gneezy and Rustichini 2000a, Nagin et al. 2002, Falk and Kosfeld 2006, Fisman and Miguel 2007, Dickinson and Villeval 2008, Boly 2011). These studies assess the direct effects of monitoring on work behavior in the monitored productivity dimension. In typical work surroundings, however, productivity is multidimensional and there are multiple ways in which workers can behave counterproductively: from showing up late, to doing sloppy work, stealing, bullying, or sabotaging other people's work; counterproductive behavior has many possible facets. Negative crowding out effects of monitoring may spill over to other productivity dimensions. These spillover effects should be incorporated when evaluating and designing monitoring and incentive schemes.

We study an experimental setup with multiple observable dimensions of productivity, in which only one dimension is monitored and incentivized. We vary (1) whether workers are monitored or not and (2) how "harsh" the incentives are. We then evaluate the effects of monitoring on the monitored dimension and on the other nonmonitored dimensions. The experimental setup we use is related to the euro currency. It is a field version of the laboratory task proposed in Belot and Schröder (2013). We recruited students to identify the provenance of euro coins. Every worker receives four boxes of coins and is asked to identify and return the coins by an appointed date. The task has the advantage of offering a menu of observable forms of counterproductive behaviors that are very common in the workplace, i.e., sloppy work, tardiness, and theft. These forms of counterproductive behavior vary in their nature and perhaps, importantly, in the nonmonetary (or moral) costs associated with them (Robinson and Bennett 1995).

Although it is obvious that sloppy work and theft affect the principal negatively, tardiness is also generally considered as undesired behavior (Robinson and Bennett 1995, Gneezy and Rustichini 2000a, Gubler et al. 2014). However, tardiness is not perceived in the same way across countries (Basu and Weibull 2003, Krupka and Weber 2013). The experiment was conducted in Germany, where there is a strong social norm of punctuality. Proper business etiquette is to be exactly on time. For example, a website targeting English speaking businessmen living in Germany (The Local, http://www.thelocal.de) ranks punctuality as the most important aspect of etiquette for doing business in Germany. Quoting, Be on time[.] Being late in Germany is a cardinal sin. Seriously. Turning up even five or ten minutes after the arranged time—especially for a first meeting—is considered personally insulting and can create a disastrous first impression. Minimize reputation damage by calling ahead with a watertight excuse if you're going to be held up." This advice is echoed on many international business websites and guides to German etiquette (see, e.g., Kwintessential, University of Frankfurt 2013). We compare three treatments with different degrees of monitoring and incentives for work quality. The first treatment (*no monitoring*) entails no monitoring at all. We contrast this treatment to treatments with monitoring and incentives. We consider two alternative monitoring and incentive schemes. The first scheme is a "low pain, low gain" incentive scheme (*monitoring and mild incentives*), which introduces a productivity target that is relatively easy to pass and a low penalty for failing to meet it. The second is a "high pain, high gain" incentive scheme (*monitoring and harsh incentives*), which introduces a difficult productivity target and a high penalty for failing to meet it. These two schemes are interesting because it is not clear a priori which of the two triggers greater effort. Harsh incentives may discipline workers and increase productivity, but incentives may also discourage the workers if the target is perceived as not worthwhile achieving. Thus, the effects of these incentive schemes on productivity are unclear ex ante.

We find evidence for negative spillover effects that appear as soon as monitoring is introduced. Specifically, we find that tardiness increases substantially: The fraction of participants who show up late increases by 35% as soon as monitoring is implemented, and the magnitude of the increase is similar independent of the incentives. Theft, on the other hand, remains constant across treatments: On average, 10% of the participants steal coins. In our experiment, the direct effect on work quality seems to be driven by incentives. We find a positive effect on work quality only when incentives are harsh. Mild incentives lead to no improvement in work quality at all, while harsh incentives reduce the number of mistakes by 40%. In a companion laboratory experiment, we replicate this result and find that the combination of productivity target and penalty is crucial to determine the effectiveness of incentives.[1]

Overall, our experimental results reveal negative spillover effects of monitoring on unmonitored productivity dimensions. The positive direct effects of monitoring seem to be contingent on harsh incentives and cannot be achieved by monitoring per se. Our results are most supportive of an interpretation related to negative reciprocity, whereby workers wish to punish the principal (for monitoring them) and do so in the least costly manner for themselves (both in monetary and nonmonetary terms).

Our results suggest that monitoring can only be efficient in combination with harsh incentives. Whether or not monitoring with harsh incentives is efficient depends on the ratio of the gains in the monitored productivity dimension to the losses in other unmonitored productivity dimensions.

The rest of this paper is structured as follows: We present the experimental design in §2 and the results and a discussion in §3. We conclude in §4.

2. Experimental Design and Procedure

We recruited students to support a research project. The task is adapted from Belot and Schröder (2013) and consists of identifying the value and country of origin of euro coins that were collected in various countries in the euro zone. Participants in our experiment had one day to complete the task from home and were requested to return the work materials at a specific deadline. Our design has several methodological advantages. It involves a job that could realistically be advertised by an economics department and that can be executed in a natural work environment, i.e., workers can take the coins home rather than working in an experimental laboratory. Additionally, we can observe multiple dimensions of productivity that arise naturally: Participants can do a poor job, be late in completing the job, or steal some of the coins. Still, it is straightforward for us to design a monitoring scheme targeting only one of these dimensions. Also, in this job, participants who failed to comply in any of these three dimensions can be categorized as behaving counterproductively, since it is possible for participants to do a perfect job, provided they are willing to do it.

We recruited student workers via a notice posted at various places on the campus of the University of Magdeburg. Interested students were asked to contact the research team by email. Those who had not participated in any previous related studies received a response email briefly explaining the task. In the

[1] In this laboratory experiment, we vary the threshold and the penalty independently. We briefly describe the design and findings in the Results section. For a detailed description, see the appendix.

email, we suggested two collection dates with the corresponding return dates and asked students to choose one of them.[2] At collection, each participant received standardized verbal instructions on how to perform the job and on the monitoring procedure.[3] After answering all open questions in a standardized way, we asked participants to indicate the exact time at which they would return the coins the next day.[4]

We contrast one treatment with no monitoring and incentives to two treatments with monitoring and incentives. In the no monitoring treatment, there is no monitoring at all and participants receive a flat fee of €20. In the two monitoring treatments, one out of the four boxes is checked. Before starting to work, participants in both monitoring treatments were informed that one out of the four boxes would be checked after returning the coins. Although we kept monitoring fixed in these two treatments, we varied the incentives associated with monitoring. In the monitoring and mild incentives treatment, participants were allowed to make 10 mistakes. If we found more than 10 mistakes in the box randomly chosen for checking, the participant would only receive €19 instead of €20. In the monitoring and harsh incentives treatment, the threshold number of mistakes was only 2. If we found more than 2 mistakes in the checked box, the participant's payment was only €5 instead of €20. The first incentive scheme is mild: It is an easy threshold to pass and the penalty is small. The second incentive is harsh: it leaves little room for mistakes and the penalty is large.[5] Note that we played on two variables at the same time to vary the incentives (threshold and penalty) and chose combinations of the two that are probably most common in the workplace. However, to get more insight into how the incentive schemes work (and affect performance in the monitored task in particular), we conducted additional treatments in a laboratory experiment that vary the penalty and the threshold independently (in a 2×2 design). We will comment more extensively on the results in the next section.

A total of 91 students participated in this study, 30 each in the no monitoring and monitoring and mild incentives treatments, and 31 in the monitoring and harsh incentives treatment. All participants were allowed to take the materials home. They received a catalog illustrating the most common euro coins and four identification tables. Each participant received a set of four boxes of euro coins collected in four different countries of the euro zone. The lid of each box indicated the country the coins were collected in. Within one set, the composition of boxes varied with respect to the value and the number of coins. Across sets, however, the composition of boxes was similar. Each participant received a total of 780 coins with a value of €114.70.

When participants returned the work materials, we wrote down the exact time the materials were returned. We also asked the participants for an estimate of the time they had worked on the task, for their field of study, and we recorded the gender. Participants in the no monitoring treatment immediately received the full payment of €20 in cash. Participants in the two monitoring treatments directly received the sure part of the payment and could collect the remaining part later (usually a day later), if they met the work quality requirements of the corresponding treatment. Participants were informed about the payment procedure before working on the task.

Compared to the no monitoring treatment, the two monitoring treatments are associated with a different payment procedure that generates some inconvenience for participants. We see this as a necessary and inherent part of introducing the monitoring technology. If we would have asked participants in the no monitoring treatment to come back a day later to collect their payment, they may have felt monitored as well. Given the nature of the task, it was impossible to run the monitoring treatments without having participants coming back. Nevertheless, we believe such inconveniences are not atypical and are often an inherent part of a monitoring scheme. In many real world examples, monitoring is indeed associated with inconveniences for the worker, e.g., monitored workers have to write extra reports, make detours in order to reach central time measurement stations, cope with delays because of quality control, or bear the discomfort of camera surveillance. Thus, we are convinced that inconveniences are a natural element of monitoring mechanisms.

When the experiment was over, we checked all returned materials with respect to coin composition and mistakes in the identification task. Whenever we observed deviations in the composition of coins, we replaced coins with identical coins or coins with similar collector's value before handing the materials to the next participant.

[2] Collection was always either Monday or Wednesday in the morning between 10:00 A.M. and 12:30 P.M. and return was the next day between 3:30 P.M. and 6:00 P.M.

[3] For a detailed overview on the written and verbal communication as well as the work materials, refer to the online appendix (available as supplemental material at http://dx.doi.org/10.1287/mnsc.2014.2089).

[4] We gave participants enough time to check their schedule for the best suitable time in the time horizon between 3:30 P.M. and 6:00 P.M. Once a participant had decided on the exact return time, we noted the time in our calendar and wrote the time on a sheet of paper that was handed to the participant.

[5] The incentive scheme was framed in a neutral language for participants. We did not use the words reward or punishment.

Table 1 **Summary Statistics**

	No monitoring (1)	Monitoring and mild incentives (2)	Monitoring and harsh incentives (3)
Work quality			
Avg. total no. of mistakes in all four boxes	10.23 (16.23)	9.97 (13.45)	6.90 (10.93)
% boxes with 0–2 mistakes (%)	76.1	71.7	83.1
% boxes with 3–10 mistakes (%)	20.6	23.3	14.4
% boxes with more than 10 mistakes (%)	3.3	5.0	2.5
Tardiness			
% participants on time (within 5 min.) (%)	56.7	33.3	35.5
% early participants (advance ≥1 min.) (%)	46.7	33.3	35.5
Median advance in min. (if early)	11 (584.90)	20 (17.04)	10 (130.31)
% late participants (delay ≥ 1 min.) (%)	13.3	43.3	45.2
Median delay in min. (if late)	4 (6.29)	5 (15.48)	8 (38.93)
Theft			
No. of participants who stole coins	3	3	3
Working time			
Avg. reported working time (in min.)	111.8 (42.58)	112.5 (45.04)	124.5 (47.69)
Penalty			
% participants eligible for full payment (%)	—	100	83.9
% participants collected full payment (%)	—	50	80.6

Note. Standard deviations in parentheses.

3. Results

3.1. Summary Statistics

Table 1 shows summary statistics for the behaviors of interest across the three treatments. Regarding the productivity in the monitored dimension first, we find that the quality of work is on average higher in the monitoring and harsh incentives treatment than in the no monitoring and monitoring and mild incentives treatments. In fact, quality in the no monitoring and the monitoring and mild incentives treatments is very similar. In these two treatments, workers make 10 mistakes on average (2.5 per box), while they make on average 7 mistakes (1.7 per box) in the monitoring and harsh incentives treatment.

Looking more in detail at the distribution of mistakes, we find that most boxes have fewer than 2 mistakes, but this share is larger in the treatment with harsh incentives (It is 76.1% in the no monitoring treatment, 71.7% in the monitoring and mild incentives treatment, and 83.1% in the monitoring and harsh incentives). Most boxes have fewer than 10 mistakes, suggesting that this threshold was indeed an easy threshold to reach (97% in the no monitoring treatment, 95% in the monitoring and mild incentives, and 98% in the monitoring and harsh incentives treatment).[6]

Turning to the other dimensions of productivity, we find that punctuality varies substantially across treatments. The percentage of participants showing up on time is much higher in the absence of monitoring. Figure 1 illustrates a histogram of the deviation from the appointed return time for the separate treatments.[7] Although only four participants in the no monitoring treatment came back late (compared to sharp punctuality), more than 40% showed up late in the two monitoring treatments. In all treatments, a substantial fraction of the participants came back too early.[8]

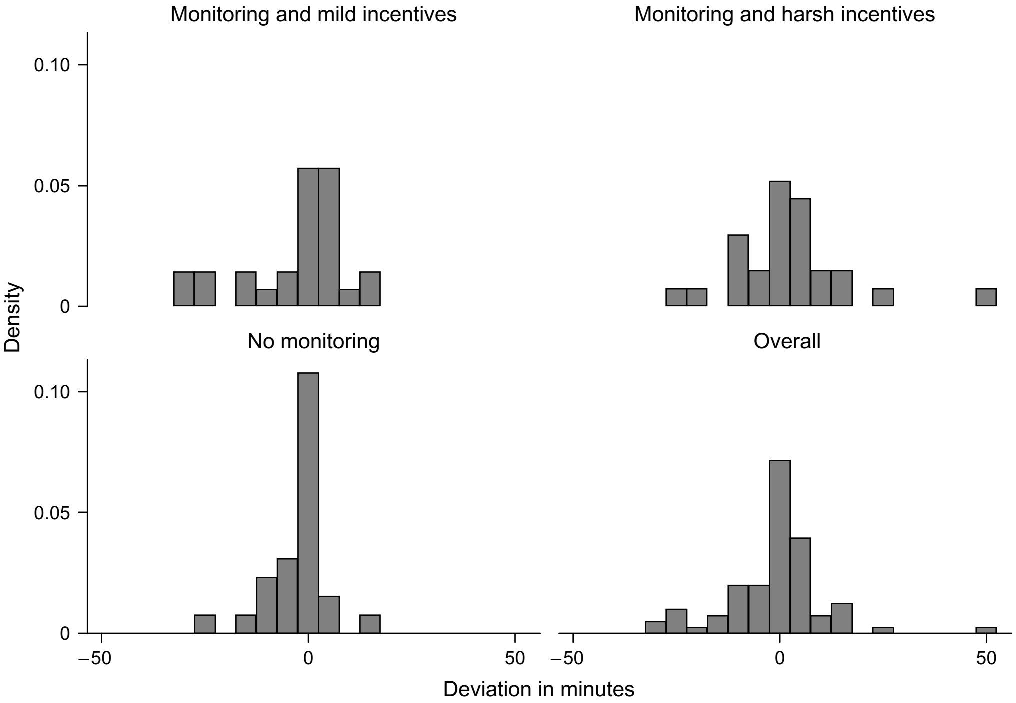
Turning to theft, we find that 10% of the participants (9 out of 91 participants) steal coins. The prevalence of theft is identical across treatments. Overall, it seems that theft in our experiment is motivated by the collectors' value of coins, rather than the nominal value

[6] In the monitoring and mild incentives treatment, all checked boxes were below the tolerated number of mistakes. Half of the participants in the monitoring and mild incentives treatment came back to collect the remaining payment. Comparing those participants who collected the remaining payment to those who did not, we do not find significant differences in the number of mistakes made (U-test, $p > 0.10$, two-tailed), stealing (Fisher Exact Test, $p > 0.10$, two-tailed), or punctuality (Fisher Exact Test, $p > 0.10$, two-tailed). In the monitoring and harsh incentives treatment, 5 participants did not meet the quality requirements. Of the 26 participants who met the requirements, 24 came back to collect the remaining payment.

[7] In the graph, we exclude outliers with a deviation above 50 minutes.

[8] It is unclear what causes participants to come back early. It could be that they try really hard not to be late and take any potentially delaying eventualities (which do not occur) into account. However, it could also be plain unpunctuality. Also, the consequences of coming back early are different to those of coming back late. By waiting, early participants can still be on time. This is clearly not the case for late participants. Most delayed participants returned the coins within the time frame. Only one participant (in the monitoring and harsh incentives treatment) returned the coins after 6:00 P.M. For early participants, we find that 15 participants (3 in the no monitoring and 6 in each monitoring treatment) returned the work materials before 3:30 P.M.

Figure 1 **Deviation from the Appointed Return Time**



of circulating coins. Participants especially steal coins that at the time of the experiment were rarely found in Germany, such as coins from the Vatican, Slovenia, or Slovakia. These are coins that have a higher collectors' value than their actual nominal value. For example, in three cases a 50¢ coin from the Vatican was stolen. On the German eBay platform this coin was sold for €3 (plus shipping) at the time of the experiment. In two cases (which occurred in different treatments), participants replaced coins from the Vatican with other coins that had the same nominal value. We categorize these acts as theft as the participants did not inform us that they replaced the coins.

Our results allow us to observe multiple dimensions of counterproductive behavior. We find that counterproductive behavior in the different dimensions is not correlated, i.e., participants who behave counterproductively in one dimension are neither more nor less likely to behave counterproductively in another dimension than other participants. Comparing individuals who steal to those who do not steal, we do not find a significant difference in tardiness (U-test, $p > 0.10$, two-tailed) or the number of mistakes (U-test, $p > 0.10$, two-tailed). Further, the number of mistakes is not correlated with the delay in minutes (Spearman Correlation, $p > 0.10$, two-tailed).

3.2. Regression Analysis

We now present a regression analysis (see Table 2) of the number of mistakes and tardiness (we do not analyze theft since there is no variation across treatments), which allows us to control for some observable characteristics of the workers. Starting with work quality, column (1) of Table 2 shows the results of a Poisson regression.[9] We find that there are 40% less mistakes under the *monitoring and harsh incentives* treatment than under no monitoring. On the other hand, we observe no significant differences between monitoring and mild incentives and no monitoring. It seems that monitoring alone does not have an effect on work quality. Work quality is only improved if monitoring is associated with harsh incentives.

Turning to punctuality, we first run a regression (column (2) of Table 2) on whether the participant showed up on time (within 5 minutes of the appointed time). We find that participants are significantly less likely to show up on time as soon as monitoring is introduced. Participants are 22% and 20% less likely to show up on time in the monitoring and mild incentives and monitoring and harsh incentives treatments, respectively. One question is whether participants show up late because they put more effort into the identification task. We asked participants how much time they spent on the task; the average reported working time was 112 minutes for the no monitoring treatment, 113 minutes for the monitoring

[9] The distribution of the number of mistakes is not normal. There is a substantial fraction of zeros and small positive values. In those cases, count data models are more appropriate. This is why we use a Poisson regression.

Table 2 Regression Analysis

	Number of mistakes (Poisson)	On time (Probit)			Early (Probit)			Late (Probit)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Monitoring and mild incentives	0.003	−0.221	−0.229	−0.243	−0.137	−0.136	−0.120	0.348	0.356	0.372
	(0.082)	(0.117)*	(0.117)*	(0.118)**	(0.119)	(0.119)	(0.123)	(0.132)***	(0.132)***	(0.143)***
Monitoring and harsh incentives	−0.407	−0.205	−0.240	−0.250	−0.105	−0.109	−0.135	0.363	0.389	0.491
	(0.089)***	(0.117)*	(0.117)**	(0.118)**	(0.120)	(0.122)	(0.122)	(0.129)***	(0.131)***	(0.136)***
Female	−0.298	−0.034	−0.047	−0.031	0.070	0.066	0.064	−0.000	0.015	0.027
	(0.074)***	(0.107)	(0.107)	(0.110)	(0.105)	(0.106)	(0.108)	(0.103)	(0.104)	(0.109)
Total mistakes	—	—	−0.007	−0.007	—	−0.001	−0.001	—	0.005	0.006
			(0.004)	(0.005)		(0.004)	(0.004)		(0.004)	(0.004)
Reported work time	—	—	0.001	0.001	—	0.000	0.000	—	0.000	0.001
			(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
Tuesday	—	—	—	−0.001	—	—	−0.166	—	—	0.231
				(0.112)			(0.108)			(0.111)**
Collection time	—	—	—	−0.074	—	—	−0.022	—	—	0.084
				(0.072)			(0.066)			(0.066)
App. return time	—	—	—	−0.024	—	—	0.059	—	—	−0.080
				(0.034)			(0.039)			(0.037)**
Constant	2.435	—	—	—	—	—	—	—	—	—
	(0.062)***									
(Pseudo) R^2	0.027	0.034	0.056	0.070	0.014	0.016	0.050	0.081	0.098	0.182
Obs.	91	91	91	91	91	91	91	91	91	91

Notes. Marginal effects are reported for Probit estimates in columns (2)–(10). Dependent variables—column (1): number of mistakes in the identification task; columns (2)–(4): dummy indicating whether the participant showed up on time (within five minutes of the appointed time); columns (5)–(7): dummy indicating whether the participant showed up early (compared to sharp punctuality); columns (8)–(10): dummy indicating whether the participant showed up late (compared to sharp punctuality).

*Significance at $p > 0.10$; **significance at $p < 0.05$; ***significance at $p < 0.01$.

and mild incentives treatment, and 124 minutes for the monitoring and harsh incentives treatment, with none of these differences being statistically significant (U-test, $p > 0.10$, two-tailed). Since the average time reported is far below 24 hours, it is unlikely that participants were under time pressure. In column (3) of Table 2, we nevertheless control whether the reported working time and the quality of work explain the differences in punctuality. In column (4) of Table 2, we additionally control for the day of the week on which participants had to return the work materials, for the time coins were collected, and for the appointed return time. The results remain unchanged when controlling for these additional variables.

The question is whether this decrease in punctuality is driven by the fact that more participants come early, or whether it is driven by more participants coming late. Columns (5)–(10) of Table 2 look at the probability of returning the work materials early or late (compared to sharp punctuality). We only find significant differences in the probability of being late. Participants are 35% and 36% more likely to be late under monitoring and mild incentives and monitoring and harsh incentives, respectively (column (8)). The effects of monitoring remain if we control for the total number of mistakes and the reported work time (columns (6) and (9)), which indicates that there is no relationship between effort in the identification task and tardiness. We also control for the day of the week, the actual collection time, and the appointed return time (columns (7) and (10)). Again, we find that participants are significantly more likely to be late in the two monitoring treatments compared to the no monitoring treatment.[10] It seems that introducing monitoring per se results in a negative spillover effect on punctuality and that these spillovers are unaffected by the level of incentives associated with monitoring.

Our experimental design varies the incentives by playing on two variables at the same time; the threshold and the penalty if this threshold is not met. Since we see a substantial increase in productivity in the identification task with harsher incentives, the question is whether this increase is driven by the higher penalty, the more difficult threshold, or both.

[10] Interestingly, we also find significant effects of the day of the week and the appointed return time on the probability of being late. Participants who return the work materials on a Tuesday are 23% more likely to be late compared to participants who return the materials on a Thursday. Further, the probability of being late decreases the later the appointed return time.

To see how these two variables affect work quality independently of each other and in combination, we conducted additional treatments in a laboratory setting where we varied the threshold and the penalty in a 2×2 design. We find that both matter: A higher penalty increases productivity and a more difficult threshold further reinforces the productivity increase when the penalty is high. Harsh incentives (difficult threshold, large penalty) appear to be the most effective way of triggering effort, while a difficult threshold with a small penalty seems to be least effective. In the latter case (difficult threshold, small penalty), incentives have an adverse effect as the number of mistakes is substantially higher than in the absence of incentives (no threshold, no penalty). We present these results in the appendix.

3.3. Discussion

We find that monitoring has a negative effect on punctuality. Independent of the level of incentives associated with monitoring, punctuality significantly decreases as soon as monitoring is introduced. What drives this crowding out effect? In the following, we will summarize some existing theories on crowding out effects and will discuss whether they can explain the observed behavior in our experiment.

One mechanism that has been proposed to explain crowding out effects is through *information*. Bénabou and Tirole (2003) argue that monitoring could negatively affect workers' perception of a task. Workers who are monitored infer that the task is difficult or unpleasant and as a consequence put less effort into the monitored task (Bénabou and Tirole 2003).

Sliwka (2007) proposes that monitoring could reveal information about peers' behavior. In his model, monitoring work quality signals that the principal expects a large fraction of workers to work sloppily. Workers, who aim at behaving in conformity with their peers, respond to this signal and choose to behave sloppily as well. It is important to note that in our task the signal is only informative for peers' behavior in the monitored productivity dimension. We showed in the results section that individuals who work sloppily are neither more nor less likely to steal or to be late. Thus, a signal on peers' work quality is not informative on their behavior in other productivity dimensions of our experiment. Both the model by Bénabou and Tirole (2003) and the model by Sliwka (2007) only predict crowding out effects on the monitored productivity dimension and cannot explain our observation that crowding out effects spill over to other productivity dimensions.

Another mechanism driving crowding out effects could be reciprocity (Rabin 1993, Frey 1993). There are multiple ways by which monitoring negatively effects workers. For a given level of effort, monitoring effectively reduces the expected payment for a worker because it is associated with a fine. Additionally, workers infer inconveniences because of the process of monitoring. Monitoring may further reduce workers' utility because of a reduction in autonomy. Reciprocal workers may want to reduce the principal's payoff as a consequence of the reduction in their own utility (Rabin 1993, Dufwenberg and Kirchsteiger 2004). It could also be that workers reciprocate distrust. Monitoring and incentives (independent of the level) may be perceived as a signal of distrust, and workers may reciprocate distrust by being less trustworthy, i.e., by caring less about the payoff of the principal (Frey 1993).

In a multidimensional context, workers should always choose the cheapest way of reciprocating. In our design, there are three ways in which workers can negatively reciprocate: (1) They can put less effort, (2) they can steal coins, and (3) they can be late in returning the work materials.[11] The first way is costly to the workers because it reduces their expected payment. The other two do not infer monetary costs for the worker (theft is even associated with monetary gains), but are associated with costs of breaking social norms. The social and the legal norm for theft is stronger than that for punctuality (e.g., Robinson and Bennett 1995). It seems reasonable to assume that tardiness is the cheapest way of reciprocating. Thus, our finding that punctuality decreases as soon as monitoring is implemented is in line with a reciprocity interpretation. It seems that workers want to retaliate for being monitored by being unpunctual.[12]

With respect to the direct effect of monitoring, we find that monitoring improves work behavior only if it is associated with harsh incentives. If the incentives associated with monitoring are mild, monitoring workers does not have any effect on the monitored productivity dimension. If the incentives are harsh, the number of mistakes falls significantly. Thus, the improvement in work quality is due to incentives rather than monitoring. In a laboratory experiment, we disentangle the effect of our two incentive components (threshold and penalty). We find that a large penalty always results in a lower number of mistakes compared to a small penalty. With respect to the threshold, we find that a difficult threshold only improves work behavior when it is associated with a large penalty. The combination of a difficult threshold

[11] All experiments were run by the researchers involved in this project. Since monitoring is not an essential part of a usual work-relation, it is clear that the monitoring choice was made by the experimenter and that tardiness would affect the experimenter.

[12] The negative effect of monitoring on workers in our experiment involves multiple dimensions, e.g., reduced expected payment, inconveniences associated with the procedure, reduced autonomy, and distrust. More research is needed to be able to disentangle the effects of the separate dimensions of monitoring on work behavior.

and a small penalty has an adverse effect on work behavior as the number of mistakes made increases substantially compared to a situation without monitoring and incentives. Our findings are in line with the existing literature on the (adverse) effects of incentives on performance (Gneezy and Rustichini 2000b, Gneezy et al. 2011) and contribute to this literature in showing that the combination of threshold and monetary incentives matters.

4. Conclusion

This paper provides field evidence on the effect of monitoring and incentives in a context where productivity is multidimensional and only one of the dimensions (work quality) is monitored. We observe negative spillovers of monitoring on unmonitored productivity dimensions. These spillover effects arise independent of the level of incentives. Thus, they appear to be driven by the mere presence of monitoring. These observed crowding out effects are in line with a model of reciprocal behavior. Workers choose to punish the principal for monitoring them, but they choose to do this through dimensions that have low costs for them.

We find that monitoring improves productivity in the monitored dimension only if it is associated with harsh incentives. Introducing monitoring and mild incentives has no effect at all on work quality. Thus, monitoring associated with mild incentives is inefficient. There is no significant improvement in work quality and tardiness increases significantly. Monitoring with harsh incentives is more effective. The number of mistakes falls substantially, but at the same time the negative spillover effects are as large as in the monitoring treatment with weak incentives.

Based on these results, we conclude that introducing a monitoring technology only pays off if (1) the incentives associated with monitoring are sufficiently harsh, (2) the dimensions that cannot be monitored either entail high moral costs or the relative gains in productivity in the monitored dimension more than compensate for the losses in other dimensions, and (3) monitoring costs for the employer are sufficiently low.

These findings relate more broadly to the literature on adverse effects of incentives (see Gneezy et al. 2011 for a recent review) and the adverse effects of control (Falk and Kosfeld 2006) and monitoring (Frey 1993). In line with this literature, we find that monitoring and mild incentives are less effective than no monitoring at all.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mnsc.2014.2089.

Acknowledgments

The authors thank Uri Gneezy, Bernd Irlenbusch, Karim Sadrieh, and three anonymous referees for valuable suggestions and comments that lead to substantial improvements. The authors also benefited from comments from participants at the European Workshop on Experimental and Behavioral Economics in Frankfurt 2013, the Royal Economic Society 2013 Conference, the 2013 Florence Workshop on Behavioural and Experimental Economics, and Seminars in Cologne and Trier. The authors thank Claudia Gorylla, Markus Hartmann, and Linh Nguyen for help in conducting the experiments. Financial support by the Institute for Fraud Prevention and the Deutsche Forschungsgemeinschaft [DFG FOR 1371] is gratefully acknowledged.

Appendix Laboratory Experiment: Threshold vs. Penalty

We conducted five additional treatments in the laboratory to find out how the threshold and the penalty affect effort in the identification task. In the laboratory experiments, we computerized the identification task and asked students to identify coins on a screen. They had to identify 204 coins that corresponded to the coins from one of the boxes in the field experiment. Since the duration of the task was shorter (50 minutes on average), we adjusted incentives to make them comparable to the field experiment and to be in accordance with expected earnings in a typical laboratory experiment.

We introduced a treatment without incentives, where participants were paid a €10 flat fee. Additionally, we ran four treatments with incentives, varying the threshold and the penalty in a 2×2 design. We offered a €10 payment to those who met the performance requirements (fewer than 2 or 10 mistakes); while those who failed would receive either €9.50 (small penalty) or €2.50 (large penalty). The five treatments are summarized in Table A.1. Note that T1 corresponds to the no monitoring treatment, T2 corresponds to the monitoring and mild incentives treatment, and T5 corresponds to the monitoring and harsh incentives treatment in the field experiment.

We ran sessions for each treatment with a between-subjects design. We had between 30 and 32 participants per treatment. Sessions were run in the Cologne Laboratory for Economic Research and subjects recruited via the online recruitment system ORSEE (Greiner 2004).

Table A.2 summarizes our results from this laboratory study. We replicate what we find in the field experiment: Mild incentives (T2) do not significantly increase effort relative to no incentives (T1) (U-test, $p = 0.17$, two-tailed). However, harsh incentives (T5) lead to significantly less mistakes than mild incentives (U-test, $p < 0.05$, two-tailed) and than no incentives at all (U-test, $p < 0.01$, two-tailed).

Table A.1 **Experimental Design and Number of Participants Laboratory Experiment**

	No threshold	Easy threshold (10 mistakes)	Difficult threshold (2 mistakes)
No penalty	T1, $N = 30$		
Small penalty (€0.50)		T2, $N = 32$	T3, $N = 32$
Large penalty (€2.50)		T4, $N = 31$	T5, $N = 32$

Table A.2 Average Number of Mistakes

	No threshold	Easy threshold (10 mistakes)	Difficult threshold (2 mistakes)
No penalty	3.7 (3.2)		
Small penalty (€0.50)		4.6 (9.8)	54.5 (77.8)
Large penalty (€7.50)		1.9 (2.8)	0.9 (1.4)

Note. Standard deviations in parentheses.

Do these effects come from the change in the threshold or the change in the penalty? We see that increasing the penalty always decreases the number of mistakes, irrespective of the threshold (U-test, $p < 0.10$, two-tailed). Making the threshold more difficult on the other hand leads to a substantial increase in the number of mistakes made when the penalty is small (U-test, $p < 0.05$, two-tailed). When the penalty is large (€7.50), a difficult threshold increases the level of effort compared to an easy threshold, but only slightly (U-test, $p < 0.10$, two-tailed).

These results show that harsh incentives increase productivity through both channels: a higher penalty increases productivity, and a more difficult threshold further reinforces the productivity increase when the penalty is high. Harsh incentives (difficult threshold, large penalty) appear to be the most effective way of triggering effort, while a difficult threshold with a small penalty seems to be least effective. In the latter case, it seems that many participants do not put much effort at all into the task (41% made more than 10 mistakes, compared to 0% in T5 (harsh incentives), 6% in T2 (mild incentives), and 3% in T1 (no incentives) and T4 (large penalty and easy threshold)).

References

Association of Certified Fraud Examiners (2012) Report to the Nations on Occupational Fraud and Abuse: 2015 global fraud study. Last accessed February 25, 2014, http://www.acfe.com/uploadedFiles/ACFE_Website/Content/rttn/2012-report-to-nations.pdf.

Basu K, Weibull JW (2003) Punctuality: A cultural trait as equilibrium. Arnott R, Greenwald B, Kanbur R, Nalebuff B, eds. *Economics for an Imperfect World: Essays in Honor of Joseph E. Stiglitz* (MIT Press, London), 163–182.

Belot M, Schröder M (2013) Sloppy work, lies and theft: A novel experimental design to study counterproductive behavior. *J. Econom. Behav. Organ.* 93:233–238.

Bénabou R, Tirole J (2003) Intrinsic and extrinsic motivation. *Rev. Econom. Stud.* 70:489–520.

Boly A (2011) On the incentive effects of monitoring: Evidence from the lab and the field. *Experiment. Econom.* 14(2):241–253.

Dickinson D, Villeval M-C (2008) Does monitoring decrease work effort? The complementary between agency and crowding-out theories. *Games Econom. Behav.* 63(1):56–76.

Dufwenberg M, Kirchsteiger G (2004) A theory of sequential reciprocity. *Games Econom. Behav.* 47(2):268–298.

Falk A, Kosfeld M (2006) The hidden costs of control. *Amer. Econom. Rev.* 96(5):1611–1630.

Fisman R, Miguel E (2007) Corruption, norms, and legal enforcement: Evidence from diplomatic parking tickets. *J. Political Econom.* 115(6):1020–1048.

Frey BS (1993) Does monitoring increase work effort? The rivalry with trust and loyalty. *Econom. Inquiry* 31(4):663–670.

Frey BS, Jegen R (2001) Motivational interactions: Effects on behavior. *Ann. Econom. Statist.* 63/64:131–153.

Gneezy U, Meier S, Rey-Biel P (2011) When and why incentives (don't) work to modify behavior. *J. Econom. Perspect.* 25(4):191–210.

Gneezy U, Rustichini A (2000a) A fine is a price. *J. Legal Stud.* 29(1):1–18.

Gneezy U, Rustichini A (2000b) Pay enough or don't pay at all. *Quart. J. Econoimcs* 115(3):791–810.

Greiner B (2004) An online recruitment system for economic experiments. Kremer K, Macho V, eds. *Forschung und wissenschaftliches Rechnen 2003* (GWDG Bericht 63, Göttingen), 73–93.

Gubler T, Larkin I, Pierce L (2014) Motivational spillovers from awards: Crowding out in a multitasking environment. HBS Working Paper 13-069, Harvard Business School, Boston.

Krupka EL, Weber RA (2013) Identifying social norms using coordination games: Why does dictator game sharing vary? *J. Eur. Econom. Assoc.* 11(3):495–524.

Kwintessential. Doing business in Germany. Accessed February 25, 2014, http://www.kwintessential.co.uk/etiquette/doing-business-germany.html.

Nagin DS, Rebitzer JB, Sanders S, Taylor LJ (2002) Monitoring, motivation, and management: The determinants of opportunistic behavior in a field experiment. *Amer. Econom. Rev.* 92(2):850–873.

Rabin M (1993) Incorporating fairness into game theory and economics. *Amer. Econom. Rev.* 83(5):1281–302.

Robinson SL, Bennett RJ (1995) A typology of deviant workplace behaviors: A multidimensional scaling study. *Acad. Management J.* 38(2):555–572.

Sliwka D (2007) Trust as a signal of a social norm and the hidden costs of incentive schemes. *Amer. Econom. Rev.* 97(3):999–1012.

The Local: Germany's news in English. Ten tips for German business etiquette. Accessed February 25, 2014, http://www.thelocal.de/galleries/news/1773.

University of Frankfurt (International Office) (2013) Guide to German culture, customs and etiquette. Accessed February 25, 2014, http://www2.uni-frankfurt.de/46329991/Guide-to-German-culture_and-etiquette.pdf.