The spillover effects of monitoring

The Spillover Effects of Monitoring: A Field Experiment

Michèle Belot*and Marina Schröder†

September 26, 2014

Abstract

We provide field experimental evidence of the effects of monitoring in a context where productivity is multi-dimensional 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 strong, but substantially reduces punctuality irrespectively of the associated incentives. Monitoring does not affect theft, with ten percent of participants stealing overall. Our findings are supportive of a reciprocity mechanism, whereby workers retaliate for being distrusted.

Keywords: counterproductive behavior, monitoring, field experiment

JEL: C93, J24, J30, M42, M52

*University of Edinburgh, School of Management, 30 Buccleuch Place, Edinburgh, EH8 9JT, UK, michele.belot@ed.ac.uk.

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

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 (see Frey 1993, Falk and Kosfeld 2006; Frey and Jegen 2014 for reviews of this literature). 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 for example 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 multi-dimensional and there are multiple ways in which workers can behave counterproductively: From showing up late to do 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 incentivised. 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 non-monitored 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 of-

fering 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 non-monetary (or moral) costs associated with them (Robinson and Bennett 1995).

While 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, Larkin and Pierce 2013). 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 (www.thelocal.de) ranks punctuality as the most important aspect of etiquette for doing business in Germany. Quoting: "*1. 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. Minimise 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.[1]

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 & 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 & harsh incentives)*, which introduces a difficult productivity target and a high penalty for failing to meet it. These two schemes are interesting

[1] See for example www2.uni-frankfurt.de/46329991/Guide-to-German-culture_and-etiquette.pdf and www.kwintessential.co.uk/etiquette/doing-business-germany.html

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 the productivity target and the penalty is crucial to determine the effectiveness of incentives.[2]

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 non-monetary 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 the paper is structured as follows: We present the experimen-

[2] 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, please see the Appendix.

tal design in Section 2 and present the results in Section 3. We discuss the interpretation of the results in Section 4 and conclude in Section 5.

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

[3] 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.

[4] For a detailed overview on the written and verbal communication as well as the work material, please refer to the online appendix.

the coins the next day.[5]

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. In the two monitoring treatments, 1 out of the 4 boxes is checked. Before starting to work, participants in both monitoring treatments were informed that 1 out of the 4 boxes would be checked after returning the coins. While we kept monitoring fixed in these two treatments, we varied the incentives associated with monitoring. In the *monitoring & 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 & harsh incentives* treatment, the threshold number of mistakes was only 2. If we found more than 2 mistakes in the checked box, the participants' 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.[6] 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 2x2 design). We will comment more extensively on the results in the next section.

Ninety one students participated in this study, 30 each in the *no monitoring* and *monitoring & mild incentives* treatments and 31 in the *monitoring & 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 4 boxes of euro coins

[5] 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.

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

collected in 4 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 particpants 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 due to quality control, or bear the discomfort of camera surveillance. Thus, we are convinced that inconveniences are a natural elemant 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.

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 & harsh incentives* treatment than in the *no monitoring* and *monitoring & mild incentives* treatments. In fact, quality in the *no monitoring* and the *monitoring & 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 & 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 & mild incentives* treatment, and 83.1% in the *monitoring & 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 & mild incentives,* and 98% in the *monitoring & harsh incentives* treatment).[7]

[7] In the *monitoring & mild* incentives treatment, all checked boxes were below the tolerated number of mistakes. Half of the participants in the *monitoring & 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 & 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.

Table 1 Summary statistics (standard deviations in brackets)

	no monitoring	monitoring & mild incentives	monitoring & harsh incentives
	(1)	(2)	(3)
Work quality			
avg. total no. of mistakes in all 4 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%
% participants too early ($\geq$1 min.)	46.6%	33.3%	35.5%
median advance in min. (if early)	11 (584.90)	20 (17.04)	10 (130.31)
% participants too late ($\geq$ 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)	111.83 (42.6)	112.5 (45.0)	124.5 (47.7)
Penalty			
% participants eligible for full payment	100%	100%	83.9%
% collected full payment if eligible	100%	50%	92.6%

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.[8] While only four participants in the *no monitoring* treatment came back late (compared to sharp punctuality), more than 40 percent showed up late in the two monitoring treatments. In all treatments, a substantial fraction of the participants came back too early.[9]

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.

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

[9] 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 (that 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 *moniting & harsh incentives* treatment) returned the coins after 6:00 p.m. For early participants, we find that 15 participants (3 in the *no monitornig* and 6 in each monitoring treatment) returned the work material before 3:30 p.m.

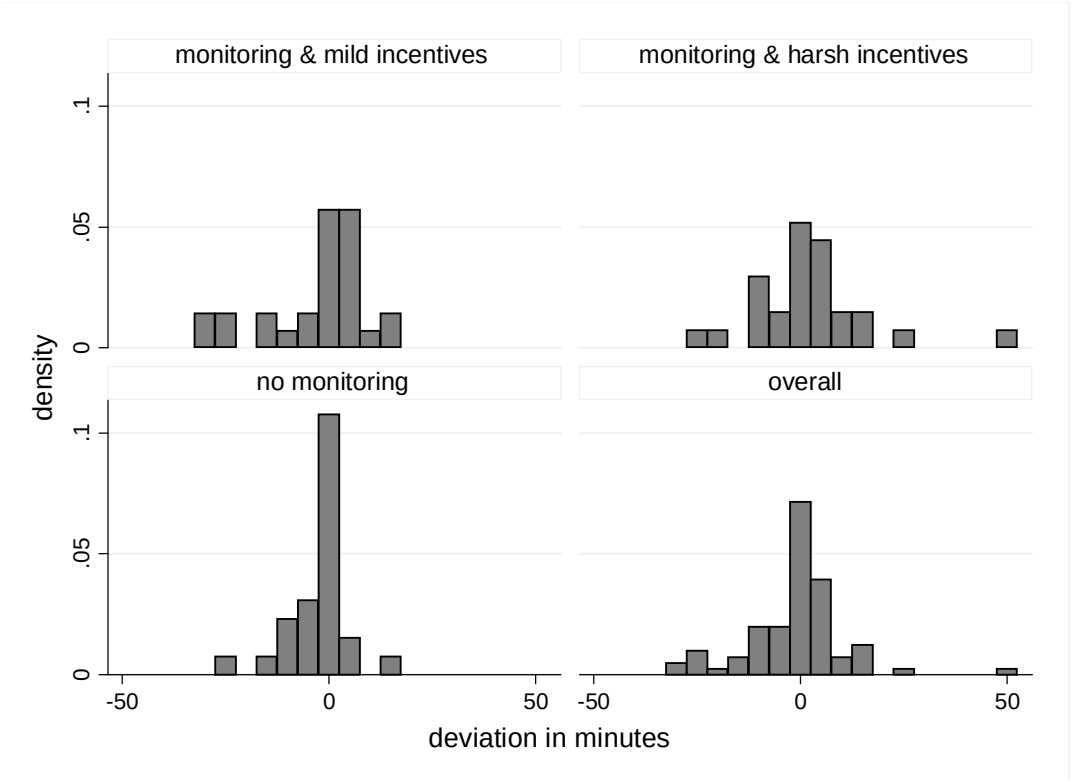

Figure 1: Deviation from the appointed return time

Overall, it seems that theft in our experiment is motivated by the collectors' value of coins, rather than the nominal value 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 cent 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 (that 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 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, Col. (1) shows the results of a Poisson regression.[10] We find that there are 40% less mistakes under the *monitoring & harsh incentives* treatment than under *no monitoring*. On the other hand, we observe no significant differences between *monitoring & 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 (Col. (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 percent less likely to show up on time in the *monitoring & mild incentives* and *monitoring & harsh incentives,* 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 and the average reported working time was 112 minutes for the *no monitoring* treatment, 113 minutes for the *monitor-*

[10] 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.

ing & mild incentives treatment, and 124 minutes for the *monitoring & harsh incentives*, 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 Col. (3) we nevertheless control whether the reported working time and the quality of work explain the differences in punctuality. In Col. (4) we additionally control for the day of the week on which participants had to return the work material, 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. Col. (5-10) look at the probability of returning the work material 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 & mild incentives* and *monitoring & harsh incentives,* respectively (Col. (8)). The effects of monitoring remain if we control for the total number of mistakes and the reported work time (Col. (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 (Col. (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.[11] 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.

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

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 & mild incentives	.003	-.221	-.229	-.243	-.137	-.136	-.120	.348	.356	.372
	(.082)	(.117)*	(.117)*	(.118)*	(.119)	(.119)	(.123)	(.132)**	(.132)***	(.143)**
monitoring & harsh incentives	-.407	-.205	-.240	-.250	-.105	-.109	-.135	.363	.389	.491
	(.089)***	(.117)*	(.117)*	(.118)**	(.120)	(.122)	(.122)	(.129)***	(.131)***	(.136)***
female	-.298	-.034	-.047	-.031	.070	.066	.064	-.000	.015	.027
	(.073)***	(.107)	(.107)	(.110)	(.105)	(.106)	(.108)	(.103)	(.104)	(.109)
total mistakes	-	-	-.007	-.007	-	-.001	-.001	-	.005	.005
			(.004)	(.005)		(.004)	(.004)		(.004)	(.004)
reported work time	-	-	.001	.001	-	.000	.000	-	.000	.001
			(.001)	(.001)		(.001)	(.001)		(.001)	(.001)
Tuesday	-	-	-	-.001	-	-	-.166	-	-	.231
				(.112)			(.108)			(.111)**
collection time	-	-	-	-.074	-	-	-.022	-	-	.084
				(.072)			(.066)			(.066)
app. return time	-	-	-	-.024	-	-	.0589	-	-	-.080
				(.034)			(.039)			(0.037)**
constant	2.435	-	-	-	-	-	-	-	-	-
	(.062)***									
pseudo) R^2	.027	.034	.056	.070	.014	.016	.050	.081	.098	.182
Obs.	91	91	91	91	91	91	91	91	91	91

*significance at p<0.10, **significance at p<0.05, ***significance at p<0.001, Marginal effects are reported for Probit estimates in Col. (2-10).
Dependent variables: Col. (1): Number of mistakes in the identification task, Col. (2-4) dummy indicating whether the participant showed up on time (within 5 minutes of the appointed time), Col. (5-7) dummy indicating whether the participant showed up early (compared to sharp punctuality), Col. (8-10) dummy indicating whether the participant showed up late (compared to sharp punctuality).

Our experimental design varies the incentives by playing on two variables at the same time: the threshold and the size of the reward for meeting the threshold. Since we see a substantial increase in the 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. 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 2x2 design. We find that both matter: a higher penalty increases productivity and a more difficult thresholdfurther 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 absense 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

conform to 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 due to the process of monitoring. Monitoring may further reduce workers' utility due to 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 trust worthy, i.e., by caring less about the payoff of the principal (Frey 1993).

In a multi-dimensional 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 material.[12] 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

[12] 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.

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.[13]

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 in the field experiment are 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 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, Meier, and Rey-Biel 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 multi-dimensional 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

[13] 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 seperate dimensions of monitoring on work behavior.

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, Meier, and Rey-Biel 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.

Appendix A Laboratory Experiment: Threshold versus 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 2x2 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 A1. Note that T1 corresponds to the "*no monitoring*" treatment , T2 corresponds to the "*monitoring & mild incentives*" treatment, and T5 corresponds to the "*monitoring & harsh incentives*" treatment in the field experiment.

Table A1 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, $N = 32$

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 ORSEE (Greiner, 2004).

Table A2 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).

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)).

Table A2: Average number of mistakes (standard deviations in brackets)

	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)

Acknowledgments

The authors thank Uri Gneezy, Bernd Irlenbusch, Karim Sadrieh, and three anonymous referees for valuable suggestions and comments that lead to substantial improvements. We 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. We 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.

References

Association of Certified Fraud Examiners. 2012. 2012 Report to the Nations on Occupational Fraud and Abuse. Available at http://www.acfe.com/uploadedFiles/ACFE_Website/Content/rttn/2012-report-to-nations.pdf, last access 25.02.2014.

Basu, K., J. W. Weibull. 2003. Punctuality: A Cultural Trait as Equilibrium. In Economics for an Imperfect World: Essays in Honor of Joseph E. Stiglitz, ed. R. Arnott, B. Greenwald, R. Kanbur, B. Nalebuff, 163–182. London: The MIT Press.

Belot, M., M. Schröder. 2013. Sloppy Work, Lies and Theft: A Novel Experimental Design to Study Counterproductive behavior. Journal of Economic Behavior and Organization 93 233-238.

Bénabou, R., J. Tirole. 2003. Intrinsic and Extrinsic Motivation. Review of Economic Studies 70 489–520.

Boly, A. 2011. On the Incentive Effects of Monitoring: Evidence from the Lab and the Field. Experimental Economics 14(2) 241–253.

Dickinson, D., M.-C. Villeval. 2008. Does Monitoring Decrease Work Effort? The Complementary Between Agency and Crowding-Out Theories. Games and Economic Behavior 63(1) 56–76.

Dufwenberg, M., G. Kirchsteiger. 2004. A Theory of Sequential Reciprocity. Games and Economic Behavior 47 268–298.

Falk, A., M. Kosfeld. 2006. The Hidden Costs of Control. American Economic Review 96(5) 1611–1630.

Fisman, R., E. Miguel. 2007. Corruption, Norms, and Legal Enforcement: Evidence from Diplomatic Parking Tickets. Journal of Political Economy 115(6) 1020–1048.

Frey, B. S. 1993. Does Monitoring Increase Work Effort? The Rivalry with Trust and Loyalty. Economic Inquiry 31(4) 663–670.

Frey, B. S., R. Jegen. 2001. Motivational Interactions: Effects on behavior. Annales of Economics and Statistics, 63/64 131–153

Gneezy, U., S. Meier, P. Rey-Biel. 2011. When and Why Incentives (Don't) Work to Modify Behavior. Journal of Economic Perspectives 25(4) 191-210.

Gneezy, U., A. Rustichini. 2000a. A Fine is a Price. Journal of Legal Studies 29(1) 1-18.

Gneezy, U., A. Rustichini. 2000b. Pay Enough or Don't Pay at All. Quarterly Journal of Econoimcs 115(3) 791–810.

Greiner, B. 2004. An Online Recruitment System for Economic Experiments. In Forschung und wissenschaftliches Rechnen 2003, ed. K. Kremer, V. Macho,73-93. GWDG Bericht 63, Göttingen.

Gubler, T., I Larkin, L. Pierce. 2013. The Dirty Laundry of Employee Award Programs: Evidence from the Field. Harvard Business School Working Paper 13-069.

Krupka, E. L., R. A. Weber. 2013. Identifying Social Norms Using Coordination Games: Why does Dictator Game Sharing Vary? Journal of the European Economic Association 11(3) 495–524.

Kwintessential. Doing Business in Germany. Available at http://www.kwintessential.co.uk/etiquette/doing-business-germany.html, last access 25.02.2014.

Nagin, D. S., J. B. Rebitzer, S. Sanders, L. J. Taylor. 2002. Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment. American Economic Review 92(2) 850-873.

Rabin, M. 1993. Incorporating Fairness into Game Theory and Economics. American Economic Review 83(5) 1281–302.

Robinson, S. L., R. J. Bennett. 1995. A Typology of Deviant Workplace Behaviors: A Multidimensional Scaling Study. Academy of Management Journal 38(2) 555–572.

Sliwka, D. 2007. Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes. American Economic Review 97(3) 999–1012.

The Local: Germany's news in English, Ten tips for German business etiquette. Available at http://www.thelocal.de/galleries/news/1773, last accesss 25.02.2014.

University of Frankfurt (International Office), 2013. Guide to German culture, customs and etiquette. Available at http://www2.uni-frankfurt.de/49378893/Guide-to-German-culture_-costums-and-etiquette-02_12_13.pdf, last access 25.02.2014.