

On the Incentive Effects of Monitoring: Evidence from the Lab and the Field

Amadou Boly*

January 2010

Abstract

Several experimental studies have showed that the crowding-out effect of monitoring may outweigh its disciplining effect, thereby reducing effort. However, most of these experiments use numeric effort tasks that agents may not be intrinsically motivated to complete. This paper reports on two similar experiments carried out in the lab and in the field using a real-effort task for which intrinsic motivation is more likely to exist. Both in the lab and in the field, we find that agency theory holds; that is, monitoring significantly increases effort on average.

JEL Codes: C9, J4, M5

Keywords: Experimental Economics, Monitoring, Crowding-Out Effect

*Research and Statistics Branch, UNIDO, Vienna International Center, P.O. Box 300, A-1400, Vienna, Austria. E-mail: a.boly@unido.org. I would like to thank Jim Engle-Warwick, Gérard Gaudet, Julie Héroux, Abraham Hollander and Claude Montmarquette. I am particularly indebted to Olivier Armantier. I would also like to thank conference participants at the ESA 2008 European Meeting in Lyon. All remaining errors are mine.

1. Introduction

Standard economic models of incentives consider that agents are likely to shirk because of effort disutility. Monitoring and sanctions may then be used to induce agents to raise their effort level in fear of punishments (see e.g. Becker 1968, Alchian and Demsetz 1972). In contrast, some behavioral models suggest that if perceived as hostile or unfair, monitoring may reduce effort by e.g. destroying intrinsic motivation (Frey 1993, Frey and Jegen 2001). Monitoring may have therefore both a disciplining and a crowding-out effect, working in opposite direction.

Several experimental studies show that the crowding-out effect of monitoring may outweigh its disciplining effect, thereby reducing effort (see e.g. Fehr and Falk 2002). Fehr and Gächter (2002) find that the crowding-out effect of imposing a fine to prevent agents from shirking may be large enough to make, on average, contracts with monitoring less efficient than contracts without monitoring. Falk and Kosfeld (2006) analyze a control mechanism in which a minimum performance level is imposed to the agent by restricting his choice set. They also find that the undermining effect of monitoring is larger than its disciplining effect.¹

The crowding-out effect of monitoring is generally explained through intrinsic motivation destruction. However, most of the existing experiments use numeric effort tasks which it may be difficult to argue that agents are intrinsically motivated to complete (Fehr and Schmidt 2007). Dickinson and Villeval (2008) propose a real-effort experiment consisting in progressing along a curve, with output measured by the height reached on the curve. They aimed at analyzing the effect of the employment relationship (distant vs. personal) on the relative importance of the disciplining and crowding-out effects of monitoring on agents' effort.² They find that the disciplining effect of monitoring dominates in both types of employment relationship, in accordance with standard theory. However, in the personal employment relationship, monitoring also reduces effort by crowding-out voluntary cooperation rather than intrinsic motivation.

In this paper, we use a framed real-effort task for which it is more likely that intrinsic motivation exists relative to chosen numeric effort. Specifically, a group of subjects (called graders) are recruited and paid by the experimenter to grade exam papers pertaining to another group of subjects (called candidates). In contrast to Dickinson and Villeval

¹The crowding-out effect of monitoring appears to extend to other contexts than the labor market, see e.g. Gneezy and Rustichini (2000), Fehr and Rockenbach (2003), or Schulze and Frank (2003).

²To implement a personal employment relationship, the authors used a partner matching protocol (i.e., same subject pairing for several rounds) and removed anonymity among partners.

(2008), our task is more comparable to one that would occur in a real-world setting.³ We conducted two sets of similar experiments in the lab in Montreal (Canada) and in the field in Ouagadougou (Burkina Faso). Contrary to the lab, subjects in the field are unaware they are participating in an experiment.

Four experimental treatments are conducted both in the lab and the field. In the Control treatment, subjects are paid a fixed amount for their grading, independent of how they perform the task. In a second treatment called “Monitoring”, we introduce a performance-related payment scheme. Specifically, graders are told that they may be imposed a monetary penalty depending on the quality of their grading. In the third treatment, we analyze the effects on effort of increasing both the monitoring rate and the monetary penalties. Finally, a gift-exchange treatment is conducted.⁴

Both in the lab and the field, we find that monitoring significantly increases the graders’ effort relative to the Control treatment, in line with agency theory. However, increasing monitoring rate and monetary penalties does not raise effort level further. Gift-exchange is found to have no effect on effort.

According to (e.g.) Falk and Kosfeld (2006), the effects of monitoring on effort may depend on initial intrinsic motivation level. To explore this question, we use the graders’ decision to reject or accept a bribe offer as a proxy for intrinsic motivation. Graders who rejected the bribe are considered motivated, while those who accepted it are considered selfish. Both in the lab and in the field, we find that monitoring has a significant disciplining effect on selfish graders, but not on motivated graders.

The remainder of this article is organized as follows. The design of the experiment is presented in section 2. Experimental results are analyzed in section 3 and section 4 concludes.

2. Experimental Design

In the following section, we describe the experimental sessions with “candidates”, followed by the lab and the field experiment with “graders”. All the experimental sessions were conducted in French and by the same experimenter.

³Grading tasks are common in the academic world and likely to be prominent to our experimental students.

⁴Strictly speaking, only the treatment in the field may be considered gift exchange as explained subsequently.

2.1. Experimental Sessions with Candidates

We conducted two typing sessions at the CIRANO's Bell Laboratory for Experimental Economics, located in Montreal (Canada). Candidates were recruited to type a dictated text using a computer. The object of these sessions was to provide us with "exam papers" to be graded in subsequent sessions. At the beginning of the session, each candidate was assigned to an isolated computer station. Instructions were read aloud, followed by questions from the candidates. In order to control the distribution of mistakes, we selected 7 out of the candidates' 23 papers.⁵ We completed the set of exam papers to 20 by making up 13 papers with various numbers of mistakes. Finally, the exam papers were only identified by a 10-character code combining digits and letters. The first two digits, going from 01 to 20, identified the order in which the graders were asked to grade the papers. Each session lasted roughly an hour, and included respectively 11 and 12 subjects.

2.2. Lab Experiment (Montreal, Canada)

The grading sessions were conducted at the CIRANO's Experimental Lab. Subjects had to grade a set of 20 papers in a precise order. The 20 papers were divided in two packs of 10. Graders were given the second pack of 10 papers only after completing the first pack, with additional written instructions to be read privately. The graders were also provided with an isolated work station, a pen, a report sheet and an answer book (i.e. a copy of the text without mistake). Instructions on how to grade the papers were read aloud, followed by questions. After spell-checking a paper, the graders had to report the number of mistakes both on the paper and on the report sheet. The sessions had no time limit but the graders could leave the lab only when their task was completed. They were informed that only a fraction of the papers had been typed by real subjects called "candidates", but this fraction was not specified. We partially explained to the lab graders how the number of mistakes they report for a paper affects a candidate's payoff. Namely, if a grader reports more than 15 mistakes, then the paper is not remunerated. In such cases, we asked the graders to check the "Fail" column on the report sheet next to the number of mistakes. If the number of reported mistakes is 15 or less, then the payoff depends on the number of mistakes. As a rule, the lower the number of mistakes reported, the higher the remuneration for a candidate. At the end of the session, subjects had to fill a short questionnaire, after which they were paid immediately in cash.

⁵We eliminated papers with too many skipped words or too many mistakes. The selection process was made to generate an appropriate distribution of mistakes and ease the graders' task.

2.3. Field Experiment (Ouagadougou, Burkina Faso)

The field experiment took place in July 2007 in Ouagadougou (the capital city of Burkina Faso) during the national exams' period.⁶ We used a local recruiting firm to hire graders who were mostly students from the University of Ouagadougou. Flyers placed around town proposed a short-term job consisting in grading exam papers. The flyers stated that the job consisted of two sessions: a grading session lasting about half a working day, followed a week later by a debriefing session during which graders would be paid. Having a high school diploma and a form of identification were the only documentation required. Interested people were asked to come register in person at the recruiting firm's location. Upon registering, graders were given the day, the time and the location of their two sessions. They were also made aware that this was a one-time job.

The grading sessions took place in a high school. Upon arrival, the subjects were gathered in a room. Instructions on how to grade the papers were read aloud, followed by questions. Each grader was then randomly assigned to a private room where he found an envelope containing 20 exam papers properly ordered, a report sheet, a red pen and an answer book. The 20 exam papers were the same as those used in the lab in Montreal. The front page of each exam paper in the field reproduced the instructions on how to grade the papers. The graders were explicitly instructed to grade the papers in the proper order. After spell-checking a paper, the graders had to report the number of mistakes both on the front page of the paper and on the report sheet. Graders were made aware that a candidate would fail the exam when more than 15 mistakes are reported. In such cases, we asked the graders to check the "Fail" column on the report sheet next to the number of mistakes. The graders were also instructed not to leave their room until they were done grading the 20 papers. We told them that we would stop by their room every 15 minutes to answer any potential question.

Once their task completed, we gave the graders an "IOU" and reminded them to come back the following week for the debriefing and payment session. In the debriefing session, graders were informed that they took part in an experiment. The nature of the experiment was explained and information was provided about the objective of the research and the use of the data collected. Finally, subjects filled a short questionnaire, after which they were paid in cash in return for the IOU.

⁶National exams must be passed to move from primary to middle school, middle to high school and high school to college. The exam period typically lasts from June to end of July.

2.4. Experimental Treatments

There are four experimental treatments. Each treatment is conducted both in the lab and in the field. In the Control treatment, subjects are paid a fixed amount for their grading, independent of how they perform the task. In the lab, the fixed amount, called a wage hereafter, was 250 Experimental Units (EU hereafter) for 20 exam papers or $12.5EU$ per paper. The conversion rate in the lab was $C\$1 = 12EU$. In the field, the wage was set at 5,000 $FCFA$ for 20 exam papers or 250 $FCFA$ per paper.

In the “Monitoring” treatment, we told each grader that we would randomly pick and control 1 of the 20 papers he graded. Then, we would calculate the absolute difference between the number of mistakes reported by the grader and the actual number of mistakes. This difference, called absolute deviation hereafter, would be used to determine the penalty to be imposed (see Table 1).⁷

In an additional monitoring treatment called “High Monitoring”, we told each grader that we would randomly pick and control 5 out of the 20 papers graded. Then, we would compute the absolute deviations. Only the worst graded paper, i.e. the one with the highest absolute deviation, would be considered to determine the monetary penalty (see Table 1). Except for the risk of being penalized, the monitoring treatments are identical to the Control treatment.

The “Gift-exchange” treatment is also identical to the Control treatment except that the wage was 40% higher (i.e. 350 EU in the lab and 7,000 $FCFA$ in the field). In the field, we implemented gift-exchange by providing a direct gift to graders. Specifically, in posters to recruit graders in the field, the announced wage was 5,000 $FCFA$. However, graders were told on the day of the experiment that the amount they would receive had been increased to 7,000 $FCFA$. This approach is similar to the surprise approach used in e.g. Gneezy and List (2006). In the lab, subjects simply received a higher wage compared to the baseline treatment. As a result, strictly speaking, the lab treatment is not gift-exchange.

In total, 180 (respectively, 248) subjects participated in the four treatments conducted in the lab (respectively, in the field). More precisely, in the lab (field), 62 (82) subjects participated in the Control treatment, 55 (82) in the Monitoring treatment, 32 (44) in the High Monitoring treatment, and 31 (40) in the Gift-exchange treatment. On average, the total earnings of a lab grader (a field grader) were $C\$31$ (6,000 $FCFA$) in the Control treatment, $C\$27.87$ (5,060.98 $FCFA$) in the Monitoring treatment, $C\$21.36$

⁷While monetary fines are not typically used in the workplace, other more common types of “fines” include verbal warnings, demotion or dismissal (Dickinson 2001).

(4,545.45*FCFA*) in the High Monitoring treatment, and C\$39.25 (8,000*FCFA*) in the Gift-exchange treatment.⁸

3. Experimental Results

In this paper, we concentrate on the 10 first papers graded in the lab and in the field.⁹ Effort is proxied by the precision of a grader, which is measured as the absolute deviation from the actual number of mistakes in a paper. The higher the absolute deviation, the lower the grader’s precision. In the next sections, we use summary statistics and Mann-Whitney tests (two-sided) to analyze the data, followed by a regression analysis. We start by the results obtained in the lab in Montreal, before turning to those obtained in the field in Ouagadougou.

3.1. Lab Results

Summary statistics on average absolute deviation in the lab in Montreal are given in Table 2, columns “Lab”. Relative to the Control treatment, graders’ precision is significantly higher at the 1% level in the Monitoring treatment (p -value = 0.001, Mann-Whitney), and at the 10% level in the High Monitoring treatment (p -value = 0.079, Mann-Whitney). Therefore, monitoring appears to increase effort on average. While graders’ precision is slightly lower in High Monitoring relative to Monitoring, the difference between these treatments is not significant (p -value = 0.432, Mann-Whitney). Finally, giving a higher wage to graders has no significant effect on effort compared to the Control treatment (p -value = 0.616, Mann-Whitney).

We now turn to a regression analysis to study treatment effects while controlling for some individual characteristics (see Table 3).¹⁰ To do so, we exploit the panel structure of the experimental data to estimate a model of the form:

$$Y_{i,t} = \beta X_{i,t} + w_{i,t} \tag{3.1}$$

⁸The Franc CFA is the currency used in Burkina Faso. In July 2007, the conversion rate was roughly C\$1 for 442.9 *FCFA*. Using the Purchasing Power Conversion data of 2007 from the UN database, we have that C\$31 received by a lab grader in the Control treatment is equivalent to 5870.50*FCFA*. A field grader in the same treatment received 6000*FCFA*. See at <http://mdgs.un.org/unsd/mdg/SeriesDetail.aspx?srid=699&crd=>

⁹In some of the experimental sessions, exam paper 11 came with a bribe offer. The grader’s decision to accept or reject the bribe, and the grading of the 10 last papers (11 to 20) are analyzed in Armantier and Boly (2008). Note that papers 01 to 10 are not affected by the presence of the bribe. Indeed, both in the lab and in the field, graders were unaware of the bribe before they reached paper 11.

¹⁰Note that female participation is low in Burkina Faso.

where $Y_{i,t}$ is the absolute difference between the number of mistakes reported by subject i for exam paper t ($t = 1, \dots, 10$) and the actual number of mistakes in exam paper t . $X_{i,t}$ is the vector of independent variables which include the subject’s age (*Age*), gender (*Female* = 1 if female, 0 otherwise), as well as experimental treatment dummies. We control for the ranking of the exam papers (*Paper Ranking*, 1 to 10) as well as interactions between the ranking and experimental treatments. Exam paper fixed effects are also included in the regression (although not reported). To control for possible grader specific effects, we model the error term as $w_{i,t} = u_i + \varepsilon_{i,t}$, where $Var(\varepsilon_{i,t}) = \sigma_\varepsilon^2$ and $Var(u_i) = \sigma_u^2$.¹¹ The model is estimated by GLS with clustered standard errors to account for correlation of unknown form among the 10 observations of each grader.¹²

The estimation results are presented in Table 4, column “Lab”. Monitoring and High Monitoring significantly increase the graders’ precision compared to the Control treatment. Yet, precision is not significantly different between Monitoring and High Monitoring.¹³ Gift-exchange has no significant effect on the graders’ effort. In addition, Table 4 indicates that women and younger subjects are more precise in their grading. The coefficients of the interaction terms between *Paper Ranking* and the monitoring treatments are positive and significant. These coefficients suggest that while monitoring increases effort level, the effects may weaken over time.

3.2. Field Results

Table 2, columns “Field”, provides summary statistics on the average absolute deviation in Ouagadougou. This table shows that the average absolute deviation is lower in the Monitoring and High Monitoring treatment compared to the Control treatment. The differences with the Control treatment are both significant at the 1% level (p -value = 0.002 for Monitoring, and p -value = 0.004 for High Monitoring, Mann-Whitney). However, High Monitoring does not significantly increase effort relative to Monitoring (p -value = 0.475, Mann-Whitney). Table 2 also shows that gift-exchange slightly increases absolute average deviation relative to the Control treatment, but the difference is not significant (p -value = 0.756, Mann-Whitney).

Table 4, column “Field”, presents regression results from the field in Ouagadougou.

¹¹Breusch-Pagan LM tests for random effects reject the null hypothesis that $\sigma_u^2 = 0$ in both regressions of Table 4.

¹²The statistical software used is Stata 10. Using two-way random effects or a count model (Poisson) does not change the results.

¹³The null hypothesis that the regression coefficients of the monitoring treatments are equal in the lab is not rejected; $\chi^2_{(1)} = 0.29$.

The dependent variable, the independent variables, and the regression model used are identical to those described in equation 3.1. As in the lab, we find that Monitoring and High Monitoring significantly increase the graders' precision compared to the Control treatment, but precision is not significantly different between Monitoring and High Monitoring.¹⁴ Gift-exchange has no significant effect on the graders' effort. We find that women are significantly more precise than men, while the subject's age has no explanatory power in the field. Again, the effect of monitoring may weaken over time as suggested by the coefficients of the interaction terms between *Paper Ranking* and the monitoring treatments.

Overall, we find similar results between the lab in Montreal and the field in Ouagadougou. This is important as external validity, which is defined as the possibility of generalizing lab results to real-life situations, has been a fundamental concern to experimental economists (see e.g. Loewenstein 1999, Starmer 1999, Levitt and List 2007).

3.3. Discussion

In the Control treatment, graders may have an incentive to grade exam papers seriously only until they find 16 mistakes in a paper. Indeed, from there on, the paper is failing "anyway". In contrast, in the monitoring treatments, graders have an incentive to find all the mistakes in an exam paper to avoid being penalized. Thus, the observed effects of monitoring may be driven by papers with 16 or more mistakes, which may have been graded seriously in the monitoring treatments but not in the Control treatment. To check the robustness of the experimental treatment effects observed previously, we divided each of the lab and field samples into two groups of exam papers: papers with 15 or less actual mistakes, and papers with more than 15 actual mistakes.¹⁵ We run a regression for each group of exam papers using equation 3.1. In the lab and in the field, the results (not reported here) show that Monitoring and High Monitoring increase precision significantly both for papers with 15 or less actual mistakes, and those with more than 15 actual mistakes.

Workers are generally heterogenous regarding their level of intrinsic motivation (see e.g. Delfgaauw and Dur 2008). According to e.g. Falk and Kosfeld (2006), in presence of monitoring, only selfish agents are likely to raise their effort level. To analyze this hypothesis, we separate motivated graders from selfish ones by using the graders' decision

¹⁴The null hypothesis that the regression coefficients of the monitoring treatments are equal in the field is not rejected; $\chi^2_{(1)} = 0.05$.

¹⁵Out of the 10 papers we are analyzing, 5 papers have 15 or less actual mistakes, and 5 have more than 15 actual mistakes.

to reject a bribe offer as a proxy for intrinsic motivation.¹⁶ Graders who rejected the bribe are considered motivated, while those who accepted the bribe are considered selfish.¹⁷ Indeed, as bribe acceptance was without negative monetary consequence either in the lab or in the field, self-maximizing behavior should have resulted in accepting the bribe offer. We divided our sample accordingly, and run a GLS regression (see equation 3.1) for each group, pooling the data from the lab and the field.

Table 5 presents the regression results which are broadly supportive of the above conjectures. Relative to the Control treatment, monitoring appears to have a significant disciplining effect on selfish graders who work significantly harder, both in the Monitoring and the High Monitoring treatment.¹⁸ In contrast, Monitoring and High Monitoring do not affect the effort level of motivated graders. They tend to work as hard with as without monitoring.¹⁹ A possible explanation may be that, with or without monitoring, motivated graders experience an increased direct disutility from breaking principles such as meritocracy and fairness in their grading task. When a majority of workers are motivated, it may therefore be possible to save on incentive schemes such as monitoring. Finally, the Gift-exchange treatment has no significant effect either on motivated graders or on selfish graders.

Both in the lab in Montreal and in the field in Ouagadougou, we find that monitoring increases effort, indicating that agency theory holds. This is in contrast to several numeric effort lab experiments reporting that the crowding-out effect of monitoring may outweigh its disciplining effect (e.g. Fehr and Gächter 2002, Falk and Kosfeld 2006). However, High Monitoring does not increase effort relative to Monitoring. This lack of difference may come from a sort of technical ceiling effect due to our real effort task; that is, beyond a given point, it is very difficult for people to improve their performance. Such an explanation is consistent with the fact that, even for selfish graders, no effort increase is observed between Monitoring and High Monitoring. Alternatively, one can also argue that the

¹⁶In the lab, bribery was introduced as follows. After completing a first pack of 10 papers, the graders were given the remaining 10 papers, with additional written instructions explaining that the author of paper 11 (a real candidate) is offering money with the following message “Please find few mistakes in my paper”. The graders were free to accept or reject the offer, and the consequences of each decision were explained. To introduce the bribe in the field, we taped with an easily removable “post-it” a banknote in paper 11. The candidate’s message (see above) was written on the post-it. For more details, see Armantier and Boly (2008).

¹⁷64.24% of graders accepted the bribe in the lab, and 46.83% accepted in the field. In total, 54.24% of graders accepted the bribe.

¹⁸The null hypothesis that the regression coefficients of the monitoring treatments are equal for selfish graders is not rejected; $\chi^2_{(1)} = 0.28$.

¹⁹For a similar result, see Nagin and al. (2002) who find that a substantial proportion of motivated employees do not respond at all to manipulations in the monitoring rates.

increase in monitoring between the two monitoring treatments was not drastic enough. The lack of additional levels of monitoring prevents establishing a robust relationship between monitoring intensity and effort level.

Gift-exchange did not increase the effort level of graders either in the lab or, more interestingly, in the field. This result lends some support to field experiments questioning the robustness of gift-exchange as an incentive device (see e.g. Gneezy and List 2006, Hennig-Schmidt, Rockenbach and Sadrieh 2008). However, Kube, Marechal and Puppe (2008) note that gift-exchange increases productivity when the gift is in kind but not monetary. Such a result suggests the need for a broader approach to how gift-exchange relationships are formed, particularly in the field (Dur 2009).

4. Conclusion

This paper analyzes the disciplining and crowding-out effect of monitoring using a lab and a field experiment. In contrast to most previous experimental studies, we use a real-effort task for which intrinsic motivation is more likely to exist. Namely, graders are recruited and paid by the experimenter to grade exam papers pertaining to candidates. Both in the lab in Montreal and in the field in Ouagadougou, we find that monitoring significantly increases effort level relative to the Control treatment, while Gift exchange does not. However, the effects of monitoring on effort hold only for a subset of subjects who were not intrinsically motivated to work hard in the first place.

References

- Alchian, A. A., and H. Demsetz (1972): "Production, Information Costs, and Economic Organization," *American Economic Review*, 62(5), 777-795.
- Armantier, O., and A. Boly (2008): "Can Corruption Be Studied in the Lab? Comparing a Field and a Lab Experiment," CIRANO Working Papers, s2008-26.
- Becker, G. S. (1968): "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, 76(2), 169-217.
- Delfgaauw, J., and R. Dur (2008): "Incentives and Workers' Motivation in the Public Sector," *Economic Journal*, 118, 171-191.
- Dickinson, D. L. (2001): "The Carrot vs. the Stick in Work Team Motivation," *Experimental Economics*, 4(1), 107-124.
- Dickinson, D. L., and M. C. Villeval (2008): "Does Monitoring Decrease Work Effort? The Complementarity between Agency and Crowding-out Theories," *Games and Economic Behavior*, 63, 56-76.
- Dur, R. (2009): "Gift Exchange in the Workplace: Money or Attention?," *Journal of the European Economic Association*, 7, 550-560.
- Falk, A., and M. Kosfeld (2006): "Distrust - The Hidden Cost of Control," *American Economic Review*, 96(5), 1611-1630.
- Fehr, E., and A. Falk (2002): "Psychological Foundations of Incentives," *European Economic Review*, 46(4-5), 687-724.
- Fehr, E., and S. Gächter (2002): "Do Incentive Contracts Crowd Out Voluntary Cooperation?," Working Paper No. 34, Institute for Empirical Research in Economics, University of Zurich.
- Fehr, E., and B. Rockenbach (2003): "Detrimental Effects of Sanctions on Human Altruism," *Nature*, 422, 137-140.
- Fehr, E., and K. M. Schmidt (2007): "Fairness and Contract Design," *Econometrica*, 75, 121-154.
- Frey, B. S. (1993): "Does Monitoring Increase Work Effort? The Rivalry with Trust and Loyalty," *Economic Inquiry*, 31, 663-670.
- Frey, B. S., and R. Jegen (2001): "Motivation Crowding Theory," *Journal of Economic Surveys*, 15(5), 589-611.
- Gneezy, U., and J. A. List (2006): "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments," *Econometrica*, 74(5), 1365-1384.

Gneezy, U., and A. Rustichini (2000): "A Fine Is a Price," *Journal of Legal Studies*, 19, 1-18.

Hennig-Schmidt, H., B. Rockenbach, and A. Sadrieh (2008): "In Search of Workers' Real Effort Reciprocity - A Field and a Laboratory Experiment," Forthcoming in *Journal of the European Economic Association*.

Kube, S., M. A. Marechal, and C. Puppe (2008): "The Currency of Reciprocity: Gift-Exchange in the Workplace," Working Paper Series, Nr. 377, University of Zürich.

Levitt, S. D., and J. A. List (2007): "What Do Laboratory Experiments Measuring Social Preferences Tell Us about the Real World," *Journal of Economic Perspectives*, 21(2), 153-174.

Loewenstein, G. (1999): "Experimental Economics from the Vantage-Point of Behavioral Economics," *Economic Journal*, 109(453), 25-34.

Nagin, D. S., J. B. Rebitzer, S. Sanders, and L. J. Taylor (2002): "Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment," *American Economic Review*, 92(4), 850-873.

Schulze, G. G., and B. Frank (2003): "Deterrence versus Intrinsic Motivation: Experimental Evidence on the Determinants of Corruptibility," *Economics of Governance*, 4, 143-160.

Starmer, C. (1999): "Experimental Economics: Hard Science Or Wasteful Tinkering?," *Economic Journal*, 109(453), 5-15.

Tables

Table 1				
Monetary Penalties				
	Monitoring Treatment		High Monitoring Treatment	
If the difference is :	Lab (in EU)	Field (in FCFA)	Lab (in EU)	Field (in FCFA)
Between 0 and 2 mistakes	0	0	0	0
Between 3 and 5 mistakes	50	1000	100	2000
Between 6 and 9 mistakes	100	2000	150	3000
10 mistakes or more	200	4000	225	4500

Table 2								
Average Absolute Deviation								
	Treatments							
	Control		Monitoring		High Monitoring		Gift Exchange	
	Lab	Field	Lab	Field	Lab	Field	Lab	Field
Mean	3.171	4.032	2.209	3.161	2.522	3.153	3.003	4.082
Std. Dev.	1.543	1.804	1.157	1.293	1.501	1.933	1.453	2.06
Min	0.3	0.8	0.2	0.8	0.3	0.5	0.4	0.6
Max	7.4	9.3	5.5	6.3	5.7	10.9	6.8	11.2
Obs.	62	82	55	82	32	43	31	39

Table 3				
Subject Pool Characteristics				
	Age		Gender	
	Lab	Field	Lab	Field
Mean	26.612	24.665	0.444	0.129
Std. Dev.	6.549	2.271	0.498	0.336
Min	18	20	0	0
Max	54	30	1	1
Obs.	178	248	180	248

Independent Variables	Lab	Field
Age	0.044 *** (0.015)	0.011 (0.048)
Gender (Female=1)	-0.713 *** (0.204)	-0.751 ** (0.332)
Monitoring Treatment	-1.305 *** (0.335)	-1.328 *** (0.321)
High Monitoring Treatment	-1.376 *** (0.361)	-1.490 *** (0.396)
Gift Exchange Treatment	-0.132 (0.475)	-0.075 (0.479)
Paper Ranking (1 to 10)	-0.043 (0.029)	-0.034 (0.025)
Paper Ranking*Monitoring Treatment	0.082 ** (0.042)	0.068 ** (0.031)
Paper Ranking*High Monitoring Treatment	0.123 *** (0.044)	0.111 *** (0.038)
Paper Ranking*Gift Exchange Treatment	0.037 (0.051)	0.004 (0.040)
Constant	1.841 *** (0.516)	3.120 *** (1.177)
Observations	1780	2478
Number of graders	178	248
σ_u	1.209	1.608

Notes: i) robust standard errors in parentheses, clustered in graders; ii) exam paper fixed effects included; iii) *** p<0.01, ** p<0.05, * p<0.1.

Independent Variables	Motivated Graders	Selfish Graders
Age	0.064 ** (0.025)	0.028 (0.021)
Gender (Female=1)	-1.058 *** (0.279)	-0.548 ** (0.251)
Monitoring Treatment	-0.416 (0.361)	-1.926 *** (0.324)
High Monitoring Treatment	-0.599 (0.371)	-2.101 *** (0.360)
Gift Exchange Treatment	0.280 (0.433)	-0.104 (0.500)
Paper Ranking (1 to 10)	0.028 (0.029)	-0.062 *** (0.024)
Paper Ranking*Monitoring Treatment	-0.021 (0.044)	0.117 *** (0.029)
Paper Ranking*High Monitoring Treatment	0.038 (0.043)	0.164 *** (0.038)
Paper Ranking*Gift Exchange Treatment	-0.018 (0.042)	0.011 (0.048)
Field	0.710 ** (0.309)	0.792 *** (0.236)
Constant	0.584 (0.795)	2.297 *** (0.610)
Observations	1620	1918
Number of graders	162	192
σ_u	1.531	1.474

Notes: i) robust standard errors in parentheses, clustered in graders; ii) exam paper fixed effects included; iii) *** p<0.01, ** p<0.05, * p<0.1.