High-Powered Performance Pay and Crowding Out of Non-Monetary Motives

David Huffman[1] Michael Bognanno[2]

April 29, 2017

Abstract

A previous literature cautions that paying workers for performance might crowd out non-monetary motives to work hard. Empirical evidence from the field, however, has been based on between-subjects designs that are best suited for detecting crowding out due to low-powered incentives. High-powered incentives in the workplace tend to increase output, but it is unknown whether this masks crowding out. This paper uses a within-subject experimental design and finds evidence that crowding out also extends to high-powered incentives, in a real work setting with paid workers. There is individual heterogeneity, however, with a minority of workers report crowding in of motivation. Thus, the impact of performance pay might depend on the mix of worker types.

Keywords: Incentives, non-cognitive skills, experiment, intrinsic motivation

JEL codes: D03, J22, J33

[1] University of Pittsburgh and IZA; e-mail: huffmand@pitt.edu

[2] Temple University and IZA; e-mail: bognannno@temple.edu

1 Introduction

A large literature in psychology, and more recently, behavioral economics, cautions that paying workers for performance might tend to crowd out non-monetary motives for doing a good job (Deci, 1971; Lepper et al., 1973; Kreps, 1997; Frey and Oberholzer-Gee, 1997; Baron and Kreps, 1999; Gneezy and Rustichini, 2000b). Many of the proposed mechanisms have in common that the agent infers something about private information of the principal, when the principal implements performance pay (see Benabou and Tirole, 2003). This could be information about enjoyability of the task, or the social value of the task, or relevance of social norms calling for hard work (Gneezy and Rustichini, 2000a and 200b; Benabou and Tirole, 2003 and 2006; Heyman and Ariely, 2004; Gneezy et al., 2011; Carpenter and Dolifka, 2014). Updating beliefs about these is, in turn, assumed to affect a marginal benefit of effort that directly enters the utility function. We use the term "non-monetary motives" to refer to this collection of potential motives, rather than "intrinsic motivation," as the latter is sometimes taken to mean task enjoyment alone (Fehr and Falk, 2002).

One source of evidence on crowding out is a large literature in psychology, which has mainly used a within-subjects experimental design: Treatment subjects temporarily experience performance pay, while control subjects do not. The key finding is that treatment subjects tend to have lower effort than control subjects, after performance pay is removed, consistent with the experience of performance pay eroding non-monetary motivations (for a meta analysis see Deci and Ryan, 1999). One limitation is that most of these studies have used student subjects in the lab, doing tasks that are not usually remunerated, such as games, drawings, or other traditionally volunteer activities.[1] Understanding the impact of performance pay on leisure or volunteer activities is clearly important,[2] but it is unclear whether results generalize to workplace settings, where tasks may not be as intrinsically

[1] This includes one of the initial crowding out experiments where subjects were students writing headlines for a student newspaper (Deci, 1971); while this comes closer to a work context, student newspapers are traditionally done on a volunteer basis and for the purposes of education and skill development. This study also involved only 8 subjects, and the measure of output was speed in completing headlines, without any assessment of quality. Jordan (1989) studied intrinsic motivation in the workplace, but without random assignment of workers to treatment and control.

[2] See, e.g., Titmuss (1970), Staw and Calder (1980), Kreps (1997), Fehr and Falk (2002), Gneezy and Rustichini (2000b), Stutzer et al. (2011)

enjoyable, and remuneration is the norm.[3] A different limitation is the potential for fatigue confounds. If treatment subjects have higher effort when performance pay is in place, they might put in less effort later on than control because they are more fatigued.[4]

Another source of evidence is a more recent literature in behavioral economics, which includes the innovation of a between-subjects design. In a seminal paper, Gneezy and Rustichini (2000b) gave different groups of subjects different levels of performance pay, or no performance pay. The main finding was that individuals with low powered performance pay did worse than volunteers, while individuals with higher powered incentives did better than volunteers.[5] They demonstrated that this is true not just in the lab, but also in the field, with individuals collecting donations for charity. This evidence is not explainable by fatigue effects, and is clearly consistent with low powered performance pay reducing non-monetary motivations.[6] A limitation is that the results on high powered performance pay remain inconclusive. It could be that increasing the strength of incentives mitigates problems of crowding out, e.g., by signaling higher task value, at the same time that it increases financial motives, yielding a "double dividend." Alternatively, high powered incentives could lead to crowding out, but the effect is concealed by the strong financial motives. One advantage of the within-subjects design is that it allows observing behavior after high powered incentives are removed, potentially revealing crowding out.

The contribution of this paper is implementing the within-subjects design from psychology for the first time in a real work setting with paid workers, and leveraging the

[3] Estevez-Sorenson and Broce (2016) provide a nice addition, conducting a field experiment with the within-subjects design, but the activity – cookie tasting and evaluation – was deliberately selected to be intrinsically enjoyable, and the subject pool consisted of students who were willing to do cookie tasting for free. Interestingly, they do not find evidence of crowding out.

[4] Some studies even find lower output for treatment in the second stage, when incentives are active, consistent with the monetary incentives being too weak to offset a reduction in non-monetary motivations (e.g., Deci, 1971). This cannot be explained with fatigue effects.

[5] Heyman and Ariely (2004) find similar results using a between subjects design, but even in the case of a task designed to have zero intrinsic task enjoyability, suggesting that information about social norms is important for crowding out.

[6] Another important paper by Gneezy and Rustichini (2000a) uses a within-subject design, but focusing on the introduction of a temporary fine. They find that imposing a small fine for picking children up late from daycare actually worsens behavior, and that there is no improvement after the fine is removed. Our paper is different because of the focus on workers. Also, we have an incentive that is powerful enough to elicit better performance while it is in place. See also Fehr and Gaechter (2002) for laboratory evidence that fines can be worse than no incentives, and Carpenter and Dolifka (2014) for lab evidence on how the impact of piece rates varies with perceived incentives of the employer. Charness and Gneezy (2011) also test the psychological effect of incentives in the health domain.

ability of this design to reveal potential crowding out by high-powered incentives.[7] The work setting was an afternoon street festival. Treatment group workers (N=20) started with a fixed wage, then experienced the addition of a high powered piece rate, and then went back to a fixed wage only. Control group workers (N=19) received a fixed wage throughout the whole festival.

Our main finding is that treatment workers had higher output than control during performance pay, but output fell over time after performance pay was removed, ultimately leading to lower output than control. The post-incentive drop in output is consistent with non-monetary motivations eroding over time after the experience of performance pay. An alternative explanation could be that treatment workers were more fatigued than control after performance pay was removed, due to higher output earlier on. We look at this possibility in a variety of ways, including asking directly about fatigue in a follow-up survey, and find little evidence to support the fatigue explanation. The survey does, on the other hand, provide direct evidence for the crowding out hypothesis, in the form of self-reports about reduced motivation. Interestingly, a minority of workers actually report crowding in of motivation, so the psychological impact of incentives may not be uniformly negative. At the end of the paper we provide some exploratory analysis on how the treatment effect is related to worker personality traits, and find that the crowding out pattern is particularly pronounced for workers who score high on the personality trait of conscientiousness.

Evidence that high powered incentives can crowd out worker's non-monetary motives has important implications for economics and for managers. It means that even though sufficiently powerful performance pay may tend to increase output in the workplace (e.g., Lazear, 2000), the size of the impact depends partly on "psychological variables" that are left out of economic models, and the impact might be even greater in the absence of negative psychological effects (Benabou and Tirole, 2003). This raises new issues for thinking about the optimal design of incentives, in terms of whether different ways of delivering incentives might have better psychological properties. The findings also add to a literature using field experiments or natural experiments to study the impact of

[7] We kept the environment as natural as possible, but the surprise introduction of temporary performance pay, necessary to reveal crowding out, is less typical for the workplace than permanent incentives. The study is thus on a continuum, between the complete artificiality of a lab experiment, and the completely normal context of a standard workplace incentive scheme.

incentives on effective labor supply,[8] adding new evidence on how workers behave in the presence and absence of temporary performance pay.[9] Temporary incentives are useful for testing crowding out, but they are also important to study in their own right; they occur quite often in practice, e.g., in the form of temporary employee sales contests.[10] The exploratory results on heterogeneity and worker personality relate to a growing literature on "non-cognitive skills" in economics.[11].

2 Design of the Experiment

2.1 Work setting

Our research uses workers that were hired by a start-up company, to promote the company's new mobile payments service at a large street festival. The start-up was willing to collaborate on research mainly because our research funds would allow hiring a larger workforce than otherwise. The start-up contracted with a marketing agency, which provides workers for promotional events. The agency advertised the job online, offering a wage of $18 per hour, and was ultimately able to hire 39 workers.[12]

The job of a worker was to talk to potential customers at the festival, and convince them to establish an entry in the start-up company's database by sending a text to the start-up. A text to the company's server automatically established a database entry linked to the individual's cell phone number. Registration in the database would allow the company to send marketing materials about its service in the future, and also provided an

[8] See Lazear (2000); Paarsch and Shearer (1999); Nagin et al. (2002); Fehr and Goette (2007); Goldberg (2013); Shi (2010); Al-Ubaydli et al. (2014) and many others.

[9] Fehr and Goette (2007), and Bellemare and Shearer study the impact of temporary changes in piece rates, but do not investigate behavior in the absence of piece rates. Delfgaauw et al. (2013) study the impact of temporarily implementing performance pay, but do not analyze behavior after the performance pay is removed. While there is some work in the business literature on the impact of temporary incentives on worker behavior, to date such analyses have not typically involved a control group, making interpretation difficult (see Lim et al., 2009).

[10] Total expenditures on such contests are estimated to have been more than 26 billion dollars in 2000 in the U.S. (Lim et al., 2009).

[11] See Heckman (2000); Bowles et al. (2001a);Heckman and Rubinstein (2001); Carneiro and Heckman (2003); Persico et al. (2004); Kuhn and Weinberger (2005); Segal (2008); Lindqvist and Westman (2011); Borghans et al., (2011); Becker et al. (2012); Segal, 2012.

[12] Our goal was to hire more workers, but this proved infeasible given limited availability of workers. The sample size is nevertheless roughly double the sample sizes of the seminal intrinsic motivation experiments in psychology (Deci, 1971). The wage of $18 an hour was typical of the wages offered by the agency.

indicator of latent demand that could potentially be used to impress venture capitalists. The service offered by the start-up allowed people to pay for products at the cash register with their phone, rather than cash or card, and having an entry in the database was also a first step for a customer to establish a payment account. There was little in the way of a quality dimension to output, because workers were given only a few details about the product, and thus had little scope to provide more or less informative sales pitches. Ex post, essentially no customers took the next step of establishing a payment account, so there is no indication that certain workers did a better job of selling the service.[13]

Our measure of individual worker output comes from the fact that the text sent by the customer included the unique ID number of the worker with whom they spoke. We observe the precise time each text arrived on the start-up company's server over the course of the festival.

The festival lasted five hours on one afternoon. It took place in a major city in the US, and extended for six blocks of a shopping street, which were closed to auto traffic. The blocks were part of a named shopping district, with a similar mix of businesses throughout. Businesses along the street, which included retailers, restaurants, and other types of service providers, operated booths during the festival that featured their products. There were also booths operated by businesses that did not have shop fronts on those particular blocks. On the day of the festival the sidewalks on both sides of the street was lined with booths, regularly spaced. The types of booths were similar from block to the next. An estimated 60,000 people visited the festival, so that there were large crowds present at all times.

The work setting was particularly attractive for studying non-monetary motives because it had two features: (1) weak monitoring of worker output by the employer; (2) accurate monitoring of output by researchers. In many work settings employers do not have perfect monitoring, which means that shirking is possible, and non-monetary motives of workers are important for output. In such settings, however, it may be even more difficult for researchers to have data on worker performance. In our study there is a "wedge" between monitoring by employer and by researchers, which comes from the fact that the start-up provided the data on sign-ups to the researchers, but not to the marketing agency that was the direct employer of the workers. This was made clear

[13] As of six weeks following the festival, only 6 customers established a payment account.

to workers at the beginning of the festival.[14] Performance payments were also made to workers by representatives of the start-up, rather than the marketing agency, so the agency could not infer worker performance from payments. The interaction of the start-up with the marketing agency was one-shot, for this particular event, and any reputation concerns of workers would have been with respect to the marketing agency and not the start-up. There were three managers for the marketing agency present during the festival, but workers could easily avoid visual monitoring by losing themselves in the huge crowd.

2.2 Treatment assignment and experiment timeline

We randomly assigned 20 workers to treatment and 19 to control on the day before the festival, stratifying on information about age, gender, and previous experience provided by the marketing agency. On the day of the festival, workers arrived early for a training session. At that point they received their laminated cards with their ID numbers. The laminated card also told the worker where in the festival they would be working, and gave them a schedule of water breaks and a longer rest break. Importantly, the workers were used to being assigned to work groups, so this should not have raised any suspicions about there being an experiment in progress.[15]

To avoid awareness that an experiment was taking place, we kept the treatment and control groups separate throughout the whole festival. We obtained a waiver of informed consent from IRB.[16] We *randomly determined beforehand* that treatment workers would be assigned to stay between 20th and 17th street while control workers would be told to stay between 17th street and 14th street. The rule about not crossing 17th street was explained as being important in order to have "equal coverage" of the festival, and was enforced by having a manager of the marketing agency posted at 17th street. During the festival, each group took water breaks, and a longer rest break, at tents at their respective ends of the festival. The treatment assignment is summarized in Figure A1 in the appendix, using a

[14] The rationale given for use of ID numbers was to allow the start-up "to track sales in different areas of the festival". When performance pay was introduced, it was made clear that the calculation of payments, and payments themselves, would come directly from the start-up. Thus, there was no contradiction between workers receiving performance pay and the agency not receiving the data on sign-ups.

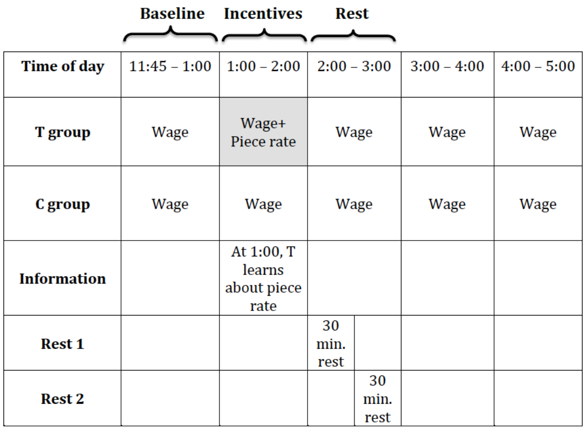
[15] The marketing agency routinely randomly assigned workers to work groups, stratified based on characteristics, to maximize "marketing effectiveness" at promotion events.

[16] The waiver was granted due to the low risk of the intervention, and the scientific value of avoiding potential Hawthorne effects.

map of the festival.

The key difference between the treatment and control conditions was not revealed to the workers initially. Figure 1 summarizes the timeline. The first time interval, denoted baseline, started at 11:45. During that roughly hour long time period, the treatment group was unaware that performance pay would be introduced.

Figure 1: Timeline for the experiment



	Baseline	Incentives	Rest			
Time of day	11:45 – 1:00	1:00 – 2:00	2:00 – 3:00		3:00 – 4:00	4:00 – 5:00
T group	Wage	Wage+ Piece rate	Wage		Wage	Wage
C group	Wage	Wage	Wage		Wage	Wage
Information		At 1:00, T learns about piece rate				
Rest 1			30 min. rest			
Rest 2				30 min. rest		

At 1:00 pm both groups checked in at their respective tents for a water break. A brief announcement was made to the treatment group by one of the experimenters: “For the next hour only there will be a special promotion from [start-up name here], where you get an extra $5 for every text that comes in with your ID number. This is on top of the $18. This only lasts for the next hour, so any text that comes in after 14:00 will not count for the $5. There won’t be any promotions later in the day.” The control group also received water, but no announcement or performance pay.

After the hour with performance pay, half of the treatment group and half of the control group went on break for 30 minutes. They rested in the shade during this time at their respective ends of the festival, eating sandwiches that were provided and being supervised by one of the research team. Subsequently these workers went back out, and the other half of treatment and control came in for a 30 minute break. After the second group was done with their break, all workers were back on the street for the remaining two hours of the festival. Neither the experimenters nor managers from the marketing agency provided any additional information, advice, or motivational speeches to workers during the rest breaks. Workers were asked to not talk on their cell phones during the festival, and adhered to this during the breaks; we cannot entirely rule out communication between workers during the working hours of the festival. The two roving managers moved throughout the whole festival, so there was no differential exposure of treatment and control workers to managers.

The geographic separation of treatment and control workers makes it relevant to discuss whether there could have been a failure of randomization, because the portion of the festival randomly assigned to treatment workers turned out to have different "customer characteristics" that mattered for the productivity of workers. Ultimately, it is an identifying assumption that there were not such differences, but we view the assumption as reasonable for several reasons. First, there were so many potential customers relative to workers (60,000 vs. 39), that availability of certain types of customers was unlikely to have been a binding constraint. Second, there was a similar mix of types of booths throughout the festival.[17] Third, the street is in the middle of the city center, so attendees came from all directions throughout the festival, which tended to ensure comparable availability of "fresh" customers at all parts of the festival; while there was one fewer side street in the treatment group area of the festival (19th street ends when it meets the principal street) there was instead a pedestrian path, so there was no important difference in accessibility to customers. Fourth, we employ a difference-in-difference analysis, using the baseline period to capture any time invariant differences in the level of customers available to workers at one end of the festival, compared to the other. Only if there were differential time trends in customer availability, for the one three block area versus the other, and only if customer availability was actually a binding constraint, would it pose a threat to identification. Finally, we can test directly for the hypothesized crowding out mechanism using self-reports about changes in non-monetary motives.

While the setting provides accurate measurement of output, it is conceivable that treatment workers found ways to "game" the performance pay incentives. For example,

[17] We obtained data on the exact booths present, and block by block location, for the 2012 festival Unfortunately we could not obtain this data for 2010, the year in which our study took place, but there is substantial overlap from year to year in vendors operating booths at the festival because many operate brick and mortar businesses in the area and participate repeatedly. There were similar although not identical numbers of booths in the treatment and control areas, 40 versus 53. There was also a similar although not identical mix of business types in both areas. The treatment area had slightly more booths focused on food and alcohol, 29 versus 24, and slightly fewer service booths, 5 versus 8. The difference in the total number of booths came mainly from retail, 4 versus 11, and the "other" category that mainly includes non-profits, with 2 booths in treatment versus 10 in control.

some could possibly have contacted friends or family via cell phone, and asked them to send texts, although this would break a rule of no cell phone use announced to workers at the beginning of the festival. This type of gaming means we might overstate the effort response of treatment workers to performance pay, if the incentives spur creativity in finding low cost ways to achieve sign ups. Our main focus, however, is on what happens to output once performance pay is removed and gaming incentives are absent.

2.3 Questionnaire

In the days following the experiment, workers were contacted to do an online questionnaire. At this point it was made clear that research was taking place, by a third party, and workers gave informed consent to participate in the survey. It was explained that the researchers would not share individual responses with the marketing agency or the start-up. Participants received $15 for participation plus any additional earnings from the trust game, described below. Out of 39 workers, 34 completed the survey.

The first part of the questionnaire asked about a well-known measure of personality from psychology, the "Big Five," which consists of five traits: conscientiousness, extraversion, agreeableness, intellect, and emotional stability.[18] The questionnaire also asked a series of other questions, about demographics and about the experience of working at the festival. Particularly important were a series of questions about fatigue, and a question about non-monetary motives. We provide the exact wordings of all questions used in the analysis as we discuss the results (wording is sometimes given in a footnote).

Respondents also participated in a modified version of the "trust game" developed by Berg et al. (1995), with other survey respondents. In this game a player can exhibit reciprocity, by choosing to "return a favor" in a one-shot anonymous interaction even though doing so entails a financial cost.[19]

[18] The questions for the Big Five are available at $http://ipip.ori.org/New\_IPIP-50-item-scale.htm$.

[19] The game involved two players, each with an endowment of $10. The first mover could choose to keep the endowment, or pass all of it to the second mover. If the money was passed, it was tripled by the experimenter. The second mover had a binary choice to keep all of the money, in the event that money was passed, or to send back $20 to the first mover. Respondents knew that it would be randomly determined who they would be matched with, among the other respondents, and which role they would play. They were asked to make a choice for both roles. We used role reversal because we use the second-mover decision as a binary measure of reciprocity. This is an incentive compatible method to elicit the choices, as either one could end up being relevant for the respondent's payoff. After survey responses were collected, the random matching of players was done. Subjects were reminded in the instructions that the agency did not know the content of the survey.

The survey allows checking for successful randomization on the basis of observable characteristics. We find no statistically significant differences between treatment and control in terms of demographics (age, gender, years of education, previous experience, English as a first language), personality type, or behavior in the trust game.[20] We also check robustness of the analysis to including worker fixed effects.

3 Behavioral Predictions

In this section we briefly discuss the predictions of three stylized models. For simplicity, the models have only two periods. In period 1 treatment workers get performance pay and control workers do not. In period 2, no-one gets performance pay. A worker chooses an effort level for each period, experiencing a cost of effort. To have non-zero effort in the absence of incentive pay, all of the models incorporate reputation concerns; putting in a minimal amount of effort helps a worker "look busy" and thus be re-hired by the marketing agency.[21] We provide formal derivations in the online appendix.

Model 1: Canonical model In a simple canonical model, treatment workers should exert more effort in period 1, compared to control workers, because the performance pay increase the marginal benefit for treatment workers. In the second period, however, treatment and control workers should choose the same effort levels, whatever is optimal in light of reputation concerns, because performance pay is gone. We assume that the performance pay earnings from one period are not large enough to change the marginal utility of income, so there is no income effect leading to lower effort in period 2.

Model 2: Non-monetary motives This model adds an extra term to the utility function in each period, θe, with θ acting as a non-monetary marginal benefit of exerting effort.

[20] Results are based on a Probit regression of the treatment dummy on all characteristics. The characteristics are not jointly significant (Chi-square; $p < 0.59$). See also Table A1 in the appendix, which provides summary statistics for treatment and control workers, as well as p-values from non-parametric tests, which show no statistically significant differences.

[21] Although we assume that treatment workers have the same reputation concerns as control workers, the experience of performance pay could conceivably have caused treatment workers to have stronger reputation concerns, because they hope they might receive performance pay in the future. This would work against finding lower output in the post-incentive period, however, and make crowding out patterns even harder to find. We also abstract away from any motive of treatment workers to strategically reduce output in period 2, in an attempt to signal the productivity benefits of performance pay, and influence the agency to use performance pay in the future. This motive was likely weak in our setting, because sign-up data were not collected by the marketing agency, and monitoring was imperfect, making signaling difficult. Also, the ability to measure individual output, a pre-requisite for performance pay, was atypical of promotion events.

The key assumption is that the presence of performance pay, or the recent experience of performance pay, affects θ. For example, a treatment worker might interpret performance pay as indicating that there is not a social norm to work hard in the absence of incentives. In the case of crowding out, performance pay makes θ smaller, thereby reducing the marginal benefit of effort in both period 1 and period 2. In this case, treatment workers have lower non-monetary motivation in period 1 and period 2 than control workers. If the financial benefits of the performance pay are large enough (high-powered), treatment workers may choose higher effort than control workers in period 1, despite crowding out. In period 2, however, treatment workers are predicted to choose lower effort than control workers. If performance pay were to cause crowding in, increasing θ, then treatment workers would have higher effort than control in both period 1 and period 2.[22]

There are alternative ways to model the interaction of performance pay with worker psychology, which would lead to similar predictions to Model (2). For example, one could assume that workers have an "output target," because they feel that they "owe" the employer a certain amount of output due to social norms to do a decent job. This would lead to similar predictions as as Model (2), but due to a subtly different mechanism: rather than reducing non-monetary motives directly, performance pay causes workers to reach psychologically-motivating goals more quickly.[23] Other psychological mechanisms, such as disappointment about removal of performance pay, were plausibly mitigated by the fact that the removal was announced at the outset. Our focus is not on disentangling subtly different psychological mechanisms that generate such predictions, but rather checking whether the basic prediction of the crowding out hypothesis is is born out in the data.

Model 3: Canonical model with fatigue effects We modify the canonical model to have fatigue effects. Fatigue is a stock that increases if previous effort is higher, and decreases if the worker rests. Specifically, the effort cost function in period 2 is a function

[22] Non (2012) and Bellemare and Shearer (2011) argue that performance pay might crowd in non-monetary motivations in the form of positive reciprocity and gift exchange.

[23] A related type of model, discussed in some previous research, involves workers in a piece rate setting having a psychological motivation (loss aversion) not to fall short of a daily "income target" (e.g., Camerer et al., 1997; Koszegi and Rabin, 2006). Our setting is different because workers have a fixed wage at all times except during the incentive hour. This means that in post-incentive hours effort did not translate into progress towards an income target, and so income target related motives could not have affected the marginal incentives to exert effort. Yet another possibility is that workers might interpret performance pay as a signal of the principal's information about the production function: How hard it is to generate a sign-up. This can lead to similar predictions to Model 2, if treatment workers update beliefs in such a way that the marginal benefit of effort is lower.

of the fatigue stock generated by effort in period 1, as well as effort in period 2. Higher effort in period 1 raises the marginal cost of effort in period 2. This model can predict higher effort of treatment than control, in period 1, due to performance pay, but lower effort than control in period 2, due to higher marginal cost of effort.

4 Results

Figure 2 shows average sign-ups per worker in the different hours of the festival. We see that, if anything, treatment workers were less productive than control in the baseline, by about 25 percent. In the hour when incentives were introduced, however, the relative position flipped, and the average treatment worker had 80 percent more sign-ups than the average treatment worker. In the hour immediately after the removal of incentives, sign-ups per worker in treatment dropped to the same level as control. Note that the figure shows double the actual sign-ups for this hour, to help correct visually for the mechanical reduction in sign-ups due to half of the workers being on break.[24] For the final two hours, sign-ups per worker were substantially lower in treatment than control, by about 75 percent and 67 percent, respectively. Differencing with respect to the respective baseline levels of sign-ups yields similar qualitative results.[25]

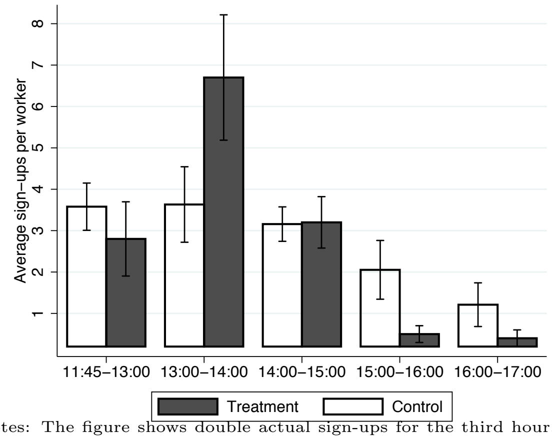
To estimate the treatment effect econometrically we use worker sign-ups aggregated to the hour; aggregating reduces noise in the effort measure and an hourly basis is natural given the structure of the experiment.[26] The number of sign-ups per worker is count data,

[24] We do not know the exact counterfactual if twice as many workers had been able to work, but doubling should give an approximate sense.

[25] For control workers, the percent changes in sign-ups (changes in log sign-ups) relative to baseline were: 0.01, -0.13, -0.56, and -1.08. For treatment workers the changes were: 0.87, 0.13, -1.72, and -1.95. The difference in differences are: 0.86, 0.26, -1.17, and -0.86 percentage points. Thus, treatment workers had a larger percent increase relative to baseline in the incentive hour, compared to control, and a larger percent decrease in later hours. In terms of levels, results are qualitatively similar. The change in the level of sign-ups relative to baseline for control workers, in each hour, were: 0.05, -0.42, -1.53, and -2.37 sign ups, respectively. The corresponding changes for treatment workers were: 3.90, 0.40, -2.30, and -2.40 sign-ups. The difference-in-differences are: 3.85, 0.82, -0.77, and -0.03 sign-ups. Thus, there was a larger increase relative to baseline level in the incentive hour, for treatment workers, and a larger decline later on, particularly in the fourth hour. The small difference in levels for the fifth hour reflects a floor effect; the level of sign-ups cannot fall below zero.

[26] One minor complication is that the baseline period was slightly longer than one hour; for the estimation, we just attribute the sign-ups generated in the brief period before festival started to the baseline hour, applying the same procedure to treatment and control. There were only 13 additional sign-ups generated in this previous time interval. The procedure tends to "inflate" performance in the baseline slightly, but it is applied to treatment and control in the same way and thus does not affect the treatment comparison. We find similar results if we instead include an additional time category for each worker,

Figure 2: Average sign-ups per worker by hour of work



Notes: The figure shows double actual sign-ups for the third hour, to correct for the mechanical reduction in sign-ups due to half of the workers being on break. Bars show +/- one standard error.

as it can only take on integer values, and must be non-negative, so it is appropriate to use count data methods. Our preferred count data estimation method is negative binomial regression, because the distribution of sign-ups involves many zero observations; a main alternative approach, Poisson regression, imposes the assumption that the mean of the dependent variable equals the standard deviation, which does not hold in the data given the skewed distribution. The negative binomial distribution is a good fit for the empirical distribution as shown in Figure A2 in the online appendix.

We estimate a difference-in-difference specification, to take into account potentially different (time invariant) differences in productivity at different ends of the festival, which would manifest as differences in baseline sign-ups for treatment versus control. The equation is as follows:

$$s_{it} = \beta + \alpha T + \gamma_2 h_2 + ... + \gamma_5 h_5 + \phi_2 h_2 \cdot T + ... + \phi_5 h_5 \cdot T + \epsilon_{it} \tag{1}$$

The dependent variable for negative binomial regression is the natural log of worker sign-ups. The constant term, and treatment dummy, are denoted β and T, respectively. The coefficients on the hours dummies, γ_2 to γ_5, give the change in log sign-ups in each hour, relative to the omitted baseline period, for control workers, i.e., approximately the percent change in average worker sign-ups. The ϕ coefficients on the interaction terms are the main coefficients of interest: They show how the percent change relative to base-

capturing the sign-ups before festival start.

line is different for treatment workers, compared to control workers, for each hour. The treatment difference is thus a difference between percent changes, measured in percentage points. We check robustness of the results to including worker fixed effects, and also to using alternative estimation methods. For all regressions, standard errors are robust and clustered at the worker level.

Table 1 reports the estimation results. Before reporting the difference-in-difference, we begin by showing the hours profiles for control and treatment workers separately, in Columns (1) and (2). The regressions are estimated using only control, and treatment workers, respectively. Column (1) shows that for control workers, sign-ups were constant initially, but then declined over the last couple of hours. Column (2) shows that for treatment workers, by contrast, there was first a strong increase relative to baseline, during the incentive hour. Over the subsequent hours, there was a substantially stronger decline relatively to baseline than for control.

Column (3) gives the difference-in-difference results. The coefficient on the treatment dummy is the difference in log sign-ups during baseline, and shows that sign-ups were roughly 20 percent lower for treatment workers than control, but the difference is not statistically significant. During the incentive hour, the change in sign-ups relative to baseline was about 86 percentage points higher for treatment than control workers and statistically significant; this reflects the 87 increase relative to baseline for treatment, versus only a 1 percent increase for control. There is no significant difference between treatment and control for the next hour, but a substantially stronger decrease in sign-ups for treatment workers, in the final two hours. The difference is 117 percentage points for the fourth hour and 87 percentage points for the fifth hour. The difference for the fourth hour is individually statistically significant ($p < 0.001$), and the coefficients for the three post-incentive hours are jointly statistically significant (Chi-square test; $p < 0.01$).

The remaining columns of Table 1 check the robustness of the treatment effect. Column (4) shows that the difference-in-difference results are similar when we add worker fixed effects, to control for any unobserved worker characteristics that might not have been successfully balanced by the randomization into treatment and control. Columns (5) and (6) check robustness to alternative estimation methods. Column (5) reports estimates from a zero-inflated Poisson regression. Poisson is robust to the incidental parameters problem that can arise in non-linear models with fixed effects, and the zero-inflated version allows

Table 1: Econometric models of the treatment effect

	(1)	(2)	(3)	(4)	(5)	(6)
13:00-14:00	0.01	0.87***	0.01	-0.13	0.04	-0.28
	(0.25)	(0.27)	(0.25)	(0.29)	(0.27)	(0.59)
14:00-15:00	-0.82***	-0.56*	-0.82***	-0.91***	-0.66**	-0.78
	(0.27)	(0.29)	(0.27)	(0.30)	(0.27)	(0.67)
15:00-16:00	-0.56	-1.72***	-0.56	-0.73**	-0.40	-1.70**
	(0.34)	(0.26)	(0.34)	(0.36)	(0.34)	(0.67)
16:00-17:00	-1.08**	-1.95***	-1.08**	-1.19***	-0.50	-2.57***
	(0.45)	(0.36)	(0.44)	(0.45)	(0.65)	(0.73)
13:00-14:00 * Treatment			0.86**	1.24***	0.76**	3.14***
			(0.36)	(0.39)	(0.39)	(1.03)
14:00-15:00 * Treatment			0.26	0.33	0.27	0.55
			(0.39)	(0.42)	(0.38)	(1.10)
15:00-16:00 * Treatment			-1.17***	-0.97**	-1.32***	-1.79*
			(0.43)	(0.48)	(0.41)	(1.08)
16:00-17:00 * Treatment			-0.86	-0.73	-1.30*	-1.39
			(0.57)	(0.61)	(0.71)	(1.24)
Treatment			-0.25			
			(0.35)			
Constant	1.28***	1.03***	1.28***			
	(0.16)	(0.31)	(0.15)			
Estimation method	N. Bin.	N. Bin.	N. Bin.	N. Bin.	Z. In. Poi..	Tobit
Worker fixed effects	No	No	No	Yes	Yes	Yes
Observations	95	100	195	195	195	195

Notes: Estimation samples are control workers only, and treatment workers only, for columns 1 and 2, respectively. The remaining columns include the full sample. The dependent variable is the log of hourly worker sign-ups, for all count data regressions, and the level of sign-ups for the Tobit regression. Interaction terms (hour*Treatment) show the difference-in-difference: How much more or less sign-ups increased relative to baseline, for Treatment vs. Control. The interaction terms are marginal effects: For negative binomial and zero-inflated Poisson (Columns 1 to 5) they give the (percentage point) difference in the percent changes relative to baseline; for Tobit they are the difference in the level of the observed dependent variable for a change in the independent variable. Regressions with worker fixed effects include a dummy variable for each worker and exclude the Treatment dummy and the constant term. Robust standard errors, clustering on workers, in parentheses. ***, **,* indicate significance at 1-, 5-, and 10-percent level, respectively.

for excess zeros.[27] Results are similar to those found using negative binomial regression. Column (6) shows results from a Tobit regression, which corrects for the mass of zeros, but does not account for the integer nature of the data (a drawback relative to count methods). In this case the dependent variable is the level of sign-ups, and we report marginal effects on the observed dependent variable.[28] The Tobit regression says that treatment workers had about 3 more sign-ups relative to baseline than control workers, during the incentive hour. During the fourth and fifth hours, the decline in sign-ups was larger for treatment

[27] Zero inflated Poisson involves a separate estimation for whether sign-ups are non-zero or not; in the absence of variables that plausibly affect having positive sign-ups, but not number of sign-ups, the equation for zero sign-ups is based on a constant term. We find similar results using regular Poisson regression, and zero-inflated negative binomial regression.

[28] We double the number of sign-ups for each worker in the fourth hour, so that the hours profile for control is corrected for the mechanical reduction in output during the break hour. This affects both treatment and control in exactly the same way, so it does not affect the estimated treatment difference from the interaction terms.

than control, by about about 1.7 sign-ups, and 1.3 sign-ups, respectively.

In summary, we find a statistically significant positive effect of incentives while they are active, and significant negative effects for the last two post-incentive hours. This pattern is consistent with the crowding out hypothesis, although it takes some time for the post-incentive drop in output to emerge. We turn below to more direct evidence on possible underlying mechanisms: fatigue, or changes in non-monetary motives.

4.1 Investigating the role of fatigue

Fatigue could potentially explain the observed treatment effect, if treatment workers are more tired than control workers, even after the rest break. Fatigue effects of performance pay are a potentially important phenomenon to study in their own right. In the questionnaire we asked workers to rate their level of fatigue at 1:00, 3:00, and 5:00 during the festival, on a five point scale, but find no indication that treatment workers were more fatigued than control at any of the three times (Mann-Whitney; $p < 0.19$, $p < 0.71$, $p < 0.36$). Figure A3 in the online appendix shows the results graphically; both treatment and control report increasing fatigue levels over the course of the festival, but there is no significant difference in the profiles over time, suggesting that workers became tired mainly due to passing hours of the day rather than the number of sign-ups achieved.

A similar result emerges from direct questions about the impact of the incentive hour on fatigue. We asked: “After having rested during the lunch break, did you still feel tired (mentally or physically) from your work during the time of the $5 bonus?” Almost all workers, 85 percent, answered either: “N.A. because I never got tired from the work I did between 1 pm and 2 pm.” or else ”The lunch break was sufficient for me to feel refreshed.” Only the remaining 15 percent chose “I was still tired from the work I did during the time of the $5 bonus (between 1 pm and 2 pm), even after the lunch break.”

As a robustness check, we re-estimated the difference-in-difference negative binomial model excluding the 15 percent of treatment workers who said that they were still fatigued from the incentive hour even after the rest break. The increase in output during the incentive hour, and the subsequent large drop in the fifth hour, are both still large and statistically significant ($p < 0.06$; $p < 0.02$). The drop for the final two hours is also still jointly significant (Chi-square test; $p < 0.06$).[29]

There are also several other findings that seem inconsistent with the predictions of our fatigue model. If fatigue effects were important, one might expect output to be lower in the first half hour following the incentive period than the second half hour; workers on the street during the first half hour have not yet had a break, whereas those in the second half hour are fresh off of a half hour break. Looking at average sign-ups for these half hour periods, we see little difference, and output is actually slightly higher in the first half hour. Specifically, average sign-ups per worker are 1.8 and 1.4 for control workers, and 1.9 and 1.3 for treatment workers, in the first and second half hour, respectively.[30] Another sign that fatigue is important would be a greater drop in output later in the festival, for individual workers who achieved an especially large number of sign-ups during the incentive hour.[31] We estimated a regression explaining hourly sign-ups in the last three hours of the festival relative to baseline, for treatment group workers. We included a dummy variable for greater than median sign-ups during the incentive hour, dummy variables for hour 3, 4, and 5, and interactions between the hour dummies and the indicator for greater than median sign-ups. The regression also included worker fixed effects. The interaction terms are all far from significant, indicating that workers with a particularly large number of sign-ups during the incentive hour did not have an especially large drop later. Thus, we do not find evidence that a higher number of individual sign-ups generated strong fatigue.

4.2 Investigating changes in non-monetary motives

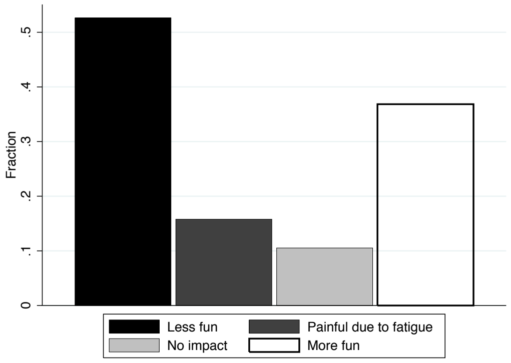
We also used the questionnaire to elicit direct evidence on whether performance pay affected non-monetary motives. We asked: "How did the experience of getting a \$5 bonus per text (between 1 and 2 pm) affect your enjoyment of the work later in the day?" Workers could give one of four responses, which were specific that the change referred to the time period when the bonus was gone.[32]

Figure 3 shows that 53 percent of the workers found working less fun in the post-

[29] The coefficients and standard errors for the interaction terms 13:00-14:00*Treatment ... 16:00-17:00*Treatment are as follows: 0.79* (0.42); 0.10 (0.38); -1.16** (0.49); -0.79 (0.63).

[30] One explanation for higher effort in the first half hour for treatment workers could be workers failing to keep track of time, and mistakenly working past 14:00 to get performance pay. In the follow-up questionnaire, however, 95 percent of treatment workers report knowing exactly when it was 14:00. Also, data on sign-ups by five minute intervals during the lunch hour does not show signs of a particularly

Figure 3: Self-reported impact of the bonus on enjoyment in post-incentive hours



incentive hours due to the experience of the bonus. Interestingly, however, about 35 percent said that the experience of performance pay actually increased enjoyment of work in the post-incentive period. Only 16 percent of workers said that they experienced fatigue, and 10 percent said the bonus had no impact on subsequent task enjoyment. The response of the majority is in line with the hypothesis that performance pay can have negative psychological effects, and is consistent with the overall lower output for treatment than control during the post-incentive ours. On the other hand, the survey responses also provide an indication that for a substantial minority, a more appropriate model might be one in which non-monetary motives are increased by performance pay.

To shed further light on potential heterogeneity, we conducted exploratory analysis on how the treatment effect varies with worker traits. The analysis is only exploratory, given the relatively small number of individuals being studied. Specifically we took three traits selected *ex ante* as possible carriers of non-monetary motives and interacted these

large number of sign-ups right after 14:00 Figure A4 in the online appendix shows these data.

[31] A particularly high number of sign-ups could also partially reflect luck rather than effort.

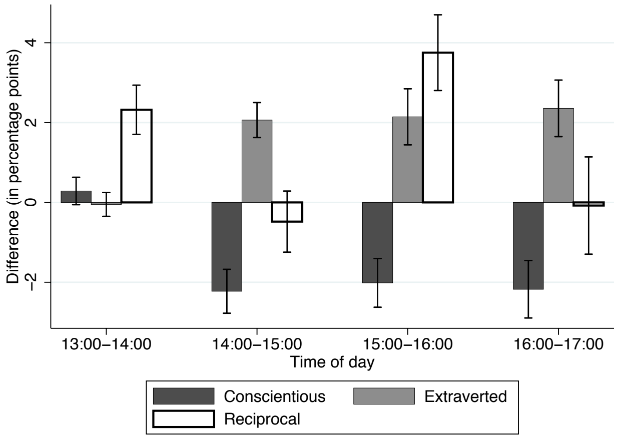
[32] The exact wording for the response categories is as follows:

1. "Getting the bonus made the work seem less fun later on, when there was no bonus."
2. "Because I worked a lot harder during the hour with the bonus, I was very tired and found working later on painful due to fatigue."
3. "Getting the bonus had no impact on my enjoyment of the work later on."
4. "Getting the bonus made the work seem more fun later on even though there was no bonus."

The question focused on motives in the post-incentive period because it is difficult to ask workers to disentangle monetary and non-monetary motives while incentives are active. We did not ask how motivation in post-incentive hours changed relative to motivation in baseline, in light of the intervening performance pay, to avoid the question becoming too complex.

with the treatment effect.[33] We selected the personality trait *conscientiousness* due to evidence that it explains working hard on laboratory tasks even in the absence of incentives, and because it tends to be correlated with positive labor market outcomes (Judge et al., 1999; Segal, 2012). We took the personality trait of *extraversion* as a potential measure of intrinsic task enjoyment, as extraverts report liking to do things like "talk to strangers at parties," which is analogous to approaching strangers at the festival.[34] We selected *positive reciprocity* due to laboratory evidence that reciprocal subjects tend to reward generous payments even if rewarding is costly for them (see Fehr and Falk, 2002). In the incentivized trust game conducted with our questionnaire, respondents had a binary choice to return a favor or not; we use this as a binary indicator of positively reciprocal tendencies.

Figure 4: Change in the treatment effect by worker traits



Negative binomial estimates. Coefficients show the change in the treatment effect for a one s.d. increase in the trait, or dichotomous change in the case of reciprocity. Error bars show +/- one standard error. Robust standard errors adjusted for clustering on worker.

Figure 4 provides an easy way to see the results. It plots coefficients showing how the treatment difference changes with a move along the dimension of a given trait (we report the regression results underlying the figure in the online appendix). Figure 4 shows no change in the treatment effect for the incentive hour, for conscientious workers, but statisti-

[33] Specifically, we estimate the following regression equation: $s_{it} = \beta + \alpha C + \gamma_2 h_2 + ... + \gamma_5 h_5 + \phi_2 h_2 \cdot T + ... + \phi_5 h_5 \cdot T + \theta_{2k} h_2 \cdot Traitk + ... + \theta_{5k} h_5 \cdot Trait_k + \alpha_{2k} h_2 \cdot T \cdot Trait_k + ... + \alpha_{5k} h_5 \cdot T \cdot Trait_k + Trait_k + \epsilon_{it}$. The α_{tk}'s are the coefficients of interest, showing whether there is a statistically significant change in the difference-in-difference at a given time t as trait k changes.

[34] The standard personality inventory we used involved ten items for each of the five personality traits. Respondents indicated how well the item described them as a person using a five point scale. The exact wording of all items is available at $http : //ipip.ori.org/New\_IPIP - 50 - item - scale.htm$.

cally significant change in the treatment difference for post-incentive hours in the negative direction, i.e., an especially strong drop in output ($p < 0.001$; $p < 0.001$; $p < 0.002$). For extraverted and reciprocal types the treatment effect is significantly different. Among reciprocal workers there was a significantly stronger positive response during the incentive period ($p < 0.001$). There was also a significant change in the treatment difference in the positive direction for the fourth hour (where the aggregate profile shows the strongest output drop; $p < 0.001$). Among more extraverted workers there was a consistent, statistically significant change of the treatment effect in the positive direction for all post-incentive hours ($p < 0.001$; $p < 0.002$; $p < 0.001$). Constructing the difference-in-difference profile separately for conscientious, extroverted, and reciprocal workers, we see that the profile for conscientious workers matches the crowding out hypothesis, whereas the other types of workers if anything exhibit crowding in. Figure A5 in the online appendix and Table A2 show the results. The results thus suggest that worker traits might be relevant for the psychological effects of incentives, with incentives potentially having a more negative effect on individuals motivated by conscientiousness, as opposed to extraversion or reciprocity.[35]

5 Conclusion

In a real work setting we implemented an experimental design from psychology, which is well suited for testing whether high-powered performance pay can crowd out workers' non-monetary motives to work hard. Treatment group workers produced more than control while performance pay was active, but exhibited progressively lower output over time, relative to control, after incentives were removed. The pattern is thus consistent with the crowding out hypothesis, although the effect takes some time to emerge. Further investigation of mechanisms provides little evidence of a role for fatigue, but does yield

[35] Changes in other personality traits – intellect, agreeableness, and emotional stability – are not associated with significant changes in the treatment effect. Other types of motives, including self-reported fatigue, self-reported reputation concerns, and self-reported "learning by doing" are also not significantly related to differences in the treatment effect. See middle and bottom panels of Figure A6 in the online appendix. Although the prediction that conscientiousness, extraversion, and reciprocity would be particularly important was formed *ex ante*, one might still be concerned that with nine traits and four coefficients for each trait some coefficients could be statistically significant just due to chance. We performed a conservative Bonferroni correction for multiple hypothesis testing and find that almost all coefficients for conscientiousness, reciprocity, and extraversion remain statistically significant (none of the coefficients for the other traits are statistically significant). Specifically, using the 36 coefficients in Figure A6 as the number of comparisons, the adjusted threshold for statistical significance is $0.05/36 = 0.001$. Two coefficients, one for conscientiousness and one for extraversion, are no longer significant but are "close" to the threshold ($p < 0.002$; $p < 0.002$).

(self-reported) evidence that performance pay affected non-monetary motives. The findings thus suggest that crowding out of non-monetary motives is not limited to low powered incentives, and that it can occur in real work settings. Future research could explore in more detail the precise channels through which crowding out occurs, using new types of treatments and survey questions.

While our findings are supportive of crowding out on average, a substantial minority of treatment workers report enhanced non-monetary motives as the result of performance pay. Also, an exploratory analysis suggests that different sources of non-monetary motives, captured by personality traits and social preferences, are associated with qualitatively different treatment effects. This heterogeneity could lead to differences in aggregate results on crowding out, across research studies using different samples. It also calls for future work on how self-selection into jobs affects the psychological impact of performance pay. For example, if the types of workers for whom performance pay does not cause crowding out tend to self select into jobs that advertise this type of compensation, there are important implications for the desirability of performance pay from both and employer and employee perspective.

References

Al-Ubaydli, O., S. Andersen, U. Gneezy, and J. A. List (2014): “Carrots that look like sticks: Toward an understanding of multitasking incentive schemes,” *Southern Economic Journal.*

Baron, J. N., and D. M. Kreps (1999): *Strategic human resources: Frameworks for general managers.* Wiley New York.

Becker, A., T. Deckers, T. Dohmen, A. Falk, and F. Kosse (2012): “The Relationship Between Economic Preferences and Psychological Personality Measures,” *Annual Review of Economics*, 4, 453–78.

Bellemare, C., and B. Shearer (2011): “On the relevance and composition of gifts within the firm: Evidence from field experiments,” *International Economic Review*, 52(3), 855–882.

Benabou, R., and J. Tirole (2003): “Intrinsic and extrinsic motivation,” *The Review of Economic Studies*, 70(3), 489–520.

——— (2006): “Incentives and Prosocial Behavior,” *The American Economic Review*, pp. 1652–1678.

Berg, J., J. Dickhaut, and K. McCabe (1995): “Trust, reciprocity, and social history,” *Games and economic behavior*, 10(1), 122–142.

Borghans, L., B. H. Golsteyn, J. Heckman, and J. E. Humphries (2011): "Identification problems in personality psychology," *Personality and Individual Differences*, 51(3), 315–320.

Bowles, S., H. Gintis, and M. Osborne (2001): "Incentive-enhancing preferences: Personality, behavior, and earnings," *The American Economic Review*, 91(2), 155–158.

Browning, M., A. Deaton, and M. Irish (1985): "A profitable approach to labor supply and commodity demands over the life-cycle," *Econometrica: Journal of the Econometric Society*, pp. 503–543.

Camerer, C., L. Babcock, G. Loewenstein, and R. Thaler (1997): "Labor supply of New York City cabdrivers: One day at a time," *The Quarterly Journal of Economics*, 112(2), 407–441.

Carneiro, P., and J. Heckman (2003): "Human Capital Policy," in *In Inequality in America: What Role for Human Capital Policy?*, ed. by A. Krueger, and J. Heckman, p. 77240, Massachusetts. MIT Press.

Carpenter, J. P., and D. Dolifka (2013): "Exploitation aversion: When financial incentives fail to motivate agents," Discussion paper, IZA Discussion Paper.

Charness, G., and U. Gneezy (2009): "Incentives to exercise," *Econometrica*, 77(3), 909–931.

Deci, E. (1971): "Effects of externally mediated rewards on intrinsic motivation," *Journal of personality and Social Psychology*, 18(1), 105–115.

Deci, E., R. Koestner, and R. Ryan (1999): "A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation.," *Psychological bulletin*, 125(6), 627.

Fehr, E., and A. Falk (2002): "Psychological foundations of incentives," *European Economic Review*, 46(4), 687–724.

Fehr, E., and L. Goette (2007): "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment," *The American Economic Review*, 97(1), 298–317.

Frey, B. S., and F. Oberholzer-Gee (1997): "The cost of price incentives: An empirical analysis of motivation crowding-out," *The American economic review*, 87(4), 746–755.

Gneezy, U., S. Meier, and P. Rey-Biel (2011): "When and why incentives (don't) work to modify behavior," *The Journal of Economic Perspectives*, pp. 191–209.

Gneezy, U., and A. Rustichini (2000a): "A Fine Is a Price," *J. Legal Stud.*, 29, 1.

——— (2000b): "Pay enough or don't pay at all," *The Quarterly Journal of Economics*, 115(3), 791–810.

Goette, L., and D. Huffman (2006): "Incentives and the Allocation of Effort Over Time: The Joint Role of Affective and Cognitive Decision Making," IZA Discussion Paper, No. 2400.

GOLDBERG, J. (2013): "Kwacha Gonna Do? Experimental Evidence about Labor Supply in Rural Malawi," *University of Maryland Working Paper.*

HECKMAN, J. (2000): "Policies to foster human capital," *Research in Economics*, 54, 3–56.

HECKMAN, J., AND Y. RUBINSTEIN (2001): "The importance of noncognitive skills: Lessons from the GED testing program," *The American Economic Review*, 91(2), 145–149.

JORDAN, P. (1986): "Effects of an extrinsic reward on intrinsic motivation: A field experiment," *The Academy of Management Journal*, 29(2), 405–412.

JUDGE, T., C. HIGGINS, C. THORESEN, AND M. BARRICK (1999): "The big five personality traits, general mental ability, and career success across the life span," *Personnel psychology*, 52(3), 621–652.

KŐSZEGI, B., AND M. RABIN (2006): "A model of reference-dependent preferences," *The Quarterly Journal of Economics*, 121(4), 1133–1165.

KREPS, D. (1997): "Intrinsic motivation and extrinsic incentives," *The American Economic Review*, 87(2), 359–364.

KUHN, P., AND C. WEINBERGER (2005): "Leadership skills and wages," *Journal of Labor Economics*, 23(3), 395–436.

LAZEAR, E. (2000): "Performance Pay and Productivity," *The American Economic Review*, 90(5), 1346–1361.

LEPPER, M., D. GREENE, AND R. NISBETT (1973): "Undermining children's intrinsic interest with extrinsic reward: A test of the" overjustification" hypothesis.," *Journal of Personality and social Psychology*, 28(1), 129.

LIM, N., M. AHEARNE, AND S. HAM (2009): "Designing sales contests: Does the prize structure matter?," *Journal of Marketing Research*, 46(3), 356–371.

LINDQVIST, E., AND R. VESTMAN (2011): "The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment," *American Economic Journal: Applied Economics*, 3(1), 101–128.

NAGIN, D., J. REBITZER, S. SANDERS, AND L. TAYLOR (2002): "Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment," *The American Economic Review*, 92(4), 850–873.

NON, A. (2012): "Gift-exchange, incentives, and heterogeneous workers," *Games and Economic Behavior*, 75(1), 319–336.

PAARSCH, H., AND B. SHEARER (1999): "The response of worker effort to piece rates: Evidence from the british columbia tree-planting industry," *Journal of Human Resources*, pp. 643–667.

PERSICO, N., A. POSTLEWAITE, AND D. SILVERMAN (2004): "The Effect of Adolescent Experience on Labor Market Outcomes: The Case of Height," *Journal of Political Economy*, 112(5).

Segal, C. (2008): "Classroom behavior," *Journal of Human Resources*, 43(4), 783–814.

Segal, C. (2012): "Working when no one is watching: Motivation, test scores, and economic success," *Management Science*, 58(8), 1438–1457.

Shi, L. (2010): "Incentive Effect of Piece-Rate Contracts: Evidence from Two Small Field Experiments," *The BE Journal of Economic Analysis & Policy*, 10(1).

Staw, B. M., B. J. Calder, R. K. Hess, and L. E. Sandelands (1980): "Intrinsic Motivation and norms about payment1," *Journal of Personality*, 48(1), 1–14.

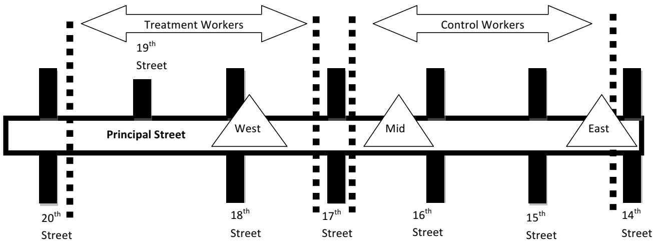
Stutzer, A., L. Goette, and M. Zehnder (2011): "Active Decisions and Prosocial Behaviour: a Field Experiment on Blood Donation," *The Economic Journal*, 121(556), F476–F493.

Titmuss, R. (1970): *The Gift Relationship: From Human Blood to Social Policy*. New Press.

A Online Appendix

A.1 Additional figures and tables

Figure A1: Treatment and control locations



Notes: Triangles represent tents operated by the start-up. Treatment and control workers took rest breaks at West and East tents, respectively.

Table A1: Sample characteristics

	Control		Treatment		Mann-Whitney
	Sample statistic	St. dev.	Sample statistic	St. dev.	p-values
Mean age	26.0	(5.49)	24.5	(3.46)	0.35
Fraction female	0.63	(0.50)	0.55	(0.51)	0.61
Fraction veteran	0.53	(0.51)	0.55	(0.51)	0.88
Fraction english second language	0.07	(0.26)	0.16	(0.37)	0.42
Fraction some college	0.80	(0.41)	0.84	(0.37)	0.75
Fraction reciprocal	0.60	(0.51)	0.37	(0.50)	0.20
Mean extraversion	0.33	(1)	-0.26	(1)	0.14
Mean agreeableness	0.26	(1)	- 0.20	(1)	0.20
Mean conscientiousness	0.01	(1)	-0.01	(1)	0.93
Mean emotional stability	-0.05	(1)	0.04	(1)	0.63
Mean intellect	0.15	(1)	-0.12	(1)	0.38

Notes: Based on 39 worker observations for Age, Female, and Veteran. Other statistics are based on the 34 survey respondents. Personality traits are standardized to have mean zero and standard deviation of 1. The final column shows p-values from nonparametric tests, of whether the trait in the corresponding row is different for control vs. treatment workers.

Figure A2: Empirical distribution vs. fitted negative binomial distribution

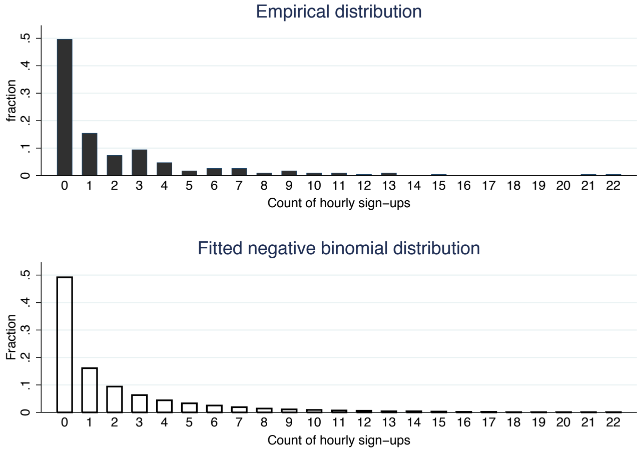
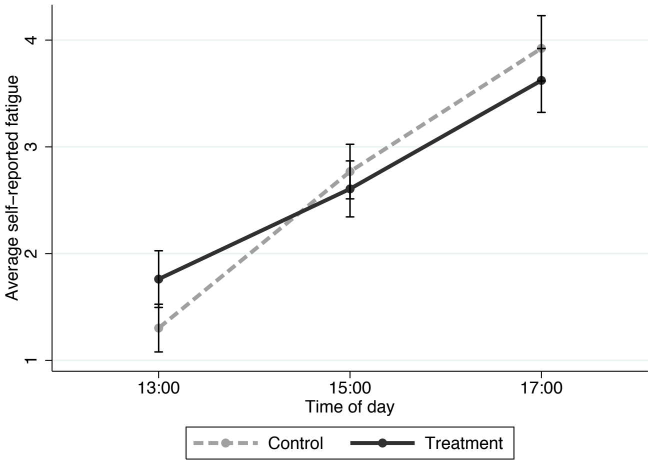
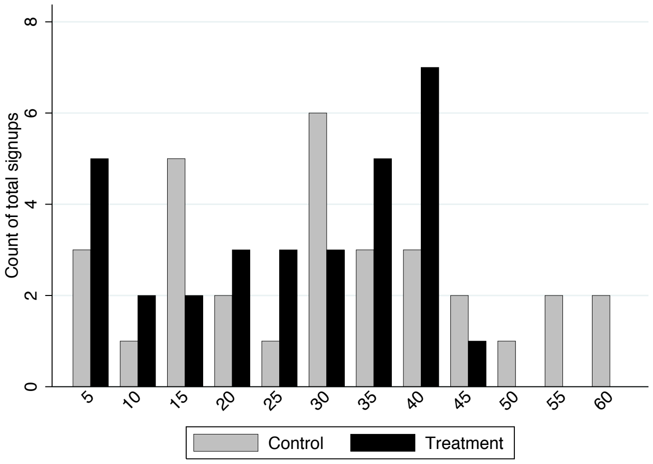


Figure A3: Self-reported fatigue



Notes: Self-reported fatigue levels at 13:00, 15:00, and 17:00, elicited in the follow up questionnaire.

Figure A4: Worker sign-ups during hour 3



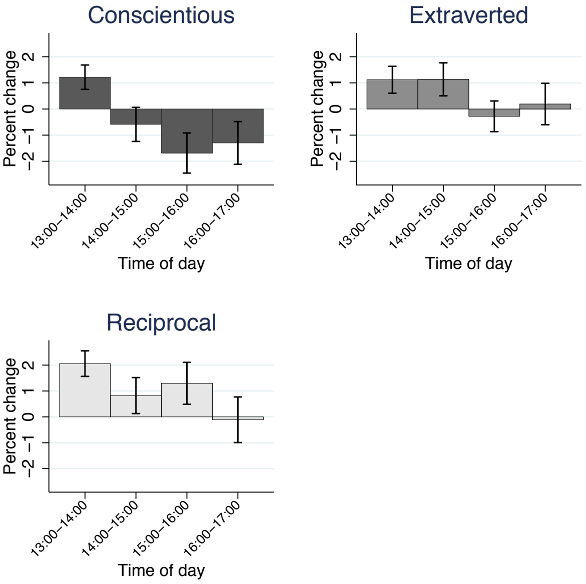
Notes: Half of treatment and control workers were on break in the first half hour, the other half in the second half hour.

Table A2: Regression estimates underlying Figure 4 and Figure A6

	(1) Conscientiousness, (2) Extraversion, (3)Reciprocity	(1) Agreeableness (2) Intellect (3) Emotional stability	(1) Reputation concerns (2) Learning by doing (3) Fatigued
13:00-14:00 * Treatment * Trait(1)	0.29	0.12	0.09
	(0.34)	(0.44)	(0.30)
14:00-15:00 * Treatment * Trait(1)	-2.33***	0.92*	-0.09
	(0.55)	(0.53)	(0.57)
15:00-16:00 * Treatment * Trait(1)	-2.02***	0.49	-0.24
	(0.61)	(0.45)	(0.56)
16:00-17:00 * Treatment * Trait(1)	-2.18***	-0.35	-0.77
	(0.72)	(0.71)	(0.53)
13:00-14:00 * Treatment * Trait(2)	-0.05	-0.44	-0.03
	(0.30)	(0.40)	(0.30)
14:00-15:00 * Treatment * Trait(2)	2.06***	-0.77	0.51
	(0.44)	(0.60)	(0.53)
15:00-16:00 * Treatment * Trait(2)	2.14***	-1.48**	-0.55
	(0.70)	(0.66)	(0.60)
16:00-17:00 * Treatment * Trait(2)	2.36***	0.83	0.60
	(0.71)	(0.80)	(0.55)
13:00-14:00 * Treatment * Trait(3)	2.32***	0.45	-0.37
	(0.62)	(0.39)	(0.36)
14:00-15:00 * Treatment * Trait(3)	-0.48	0.01	0.53
	(0.77)	(0.45)	(0.49)
15:00-16:00 * Treatment * Trait(3)	3.75***	1.08	0.05
	(0.95)	(0.85)	(0.48)
16:00-17:00 * Treatment * Trait(3)	-0.08	-0.31	0.40
	(1.22)	(0.78)	(0.77)
Other rhs. variables, coefficients supressed:			
13:00-14:00 … 16:00-15:00			
13:00-14:00 * Treatment … 16:00-15:00 * Treatment			
13:00-14:00 * Conscientiousness … 16:00-15:00 * Conscientiousness			
13:00-14:00 * Extraversion… 16:00-15:00 * Extraversion			
13:00-14:00 * Reciprocal … 16:00-15:00 * Reciprocal			
Conscientiousness			
Extraversion			
Reciprocal			
Constant			
Estimation method	N. Bin.	N. Bin.	N. Bin.
Observations	170	170	170

Notes: Negative binomial estimates. All traits are standardized so that a coefficient gives the change in the treatment effect for a one standard deviation change in the trait, with the exception of reciprocity which gives the dichotomous change. The sample is all workers who responded to the questionnaire. Robust standard errors, clustering on worker, in parentheses. ***, **,* indicate significance at 1-, 5-, and 10-percent level, respectively.

Figure A5: Treatment effect by source of non-monetary motivation



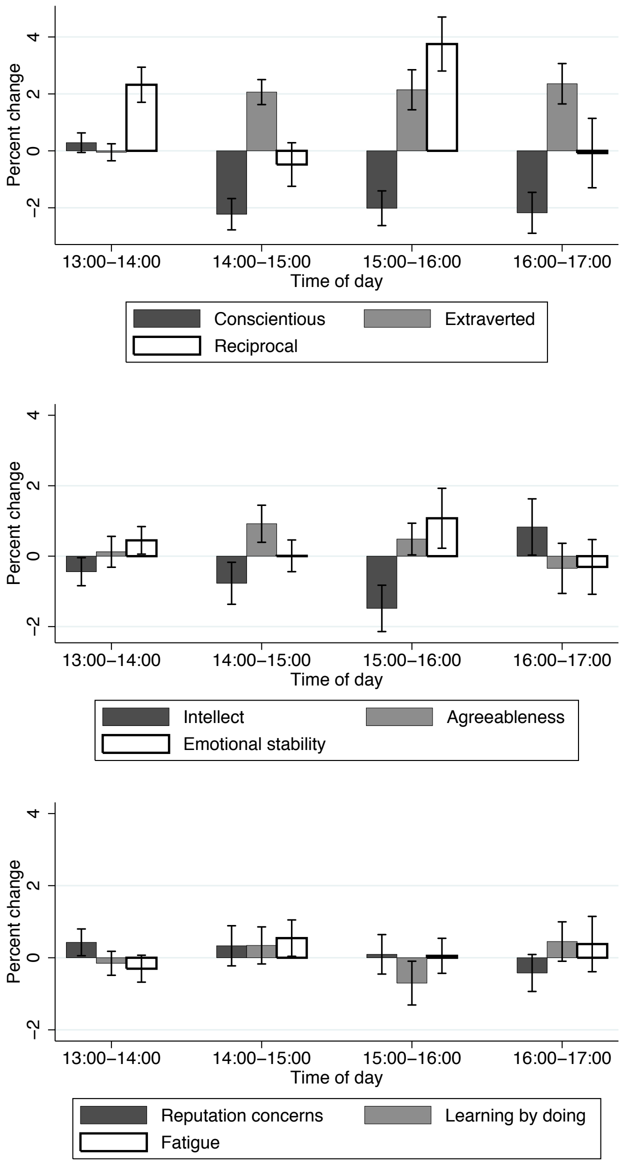
Notes: Negative binomial estimates. Coefficients show the normalized treatment effect associated with a one standard deviation increase in the trait, or the binary change in the case of reciprocity. Error bars show +/- one standard error. Robust s.e. adjusted for clustering on worker.

Table A3: Regression estimates underlying Figure A5

		Conscientiousness	Extraversion	Reciprocal
		(1)	(2)	(3)
α_2	13:00-14:00 * Treatment * Trait	0.36	0.24	2.32***
		(0.30)	(0.30)	(0.55)
α_3	14:00-15:00 * Treatment * Trait	-0.83*	0.80**	0.85
		(0.49)	(0.39)	(0.92)
α_4	15:00-16:00 * Treatment * Trait	-0.61	0.63	3.69***
		(0.48)	(0.49)	(1.26)
α_5	16:00-17:00 * Treatment * Trait	-0.73	0.70	0.61
		(0.58)	(0.59)	(1.27)
ϕ_2	13:00-14:00 * Treatment	0.86**	0.87**	-0.26
		(0.37)	(0.37)	(0.46)
ϕ_3	14:00-15:00 * Treatment	0.24	0.33	-0.03
		(0.41)	(0.45)	(0.59)
ϕ_4	15:00-16:00 * Treatment	-1.08**	-0.91*	-2.39***
		(0.48)	(0.53)	(0.87)
ϕ_5	16:00-17:00 * Treatment	-0.57	-0.50	-0.72
		(0.53)	(0.55)	(0.80)
	Other rhs. variables, coefficients supressed:			
	13:00-14:00 . . . 16:00-15:00			
	13:00-14:00 * Trait . . . 16:00-15:00 * Trait			
	Trait			
	Constant			
	Estimation method	N. Bin.	N. Bin.	N. Bin.
	Observations	170	170	170
	Normalized treatment effects plotted in Figure 5			
	$\alpha_2 + \phi_2$	1.22***	1.12**	2.06***
		(0.47)	(0.51)	(0.49)
	$\alpha_3 + \phi_3$	-0.59	1.14*	0.82
		(0.65)	(0.63)	(0.70)
	$\alpha_4 + \phi_4$	-1.68**	-0.28	1.30
		(0.77)	(0.59)	(0.81)
	$\alpha_5 + \phi_5$	-1.30	0.19	-0.11
		(0.82)	(0.79)	(0.88)

Notes: Negative binomial estimates. Traits for Columns (1) to (3) are conscientiousness, extraversion, and reciprocity, respectively. $\alpha_t + \phi_t$ gives the normalized treatment effect for workers who score high on a given trait. All traits are standardized so that a coefficient gives the change in the treatment effect for a one standard deviation change, with the exception of reciprocity which gives the dichotomous change. Standard errors are calculated for a sum of random variables. Statistical significance is based on Chi-squared test. The sample is all workers who responded to the questionnaire. Robust standard errors, clustering on worker, are in parentheses. ***, **,* indicate significance at 1-, 5-, and 10-percent level, respectively.

Figure A6: Change in the treatment effect by worker traits



Notes: The top panel reproduces Figure 4 in the text. Negative binomial estimates. Coefficients show the impact on the normalized treatment effect of a one s.d. increase in the trait, or dichotomous change in the case of reciprocity. The normalized treatment effect is the effect of the treatment differencing by the baseline difference in output. The questionnaire measured the Big Five personality traits, as well as self-reported reputation concerns, learning by doing, and fatigue as of 15:00 (footnotes in text give exact wordings). Error bars show +/- one standard error. Robust standard errors, clustering on worker, in parentheses.

A.2 Models and proofs

A.0.1 Models

For simplicity all models have just two periods, and workers maximize utility by choosing effort levels for each period.[36] In period 1 control and treatment workers get a base wage w, but treatment workers also have a performance pay rate of z. In period 2 both groups of workers just get w. In all models workers have convex costs of effort in each period, $c_t(e_t)$, but we introduce a more complex effort cost function with fatigue spillovers in Model (3). Although reputation concerns were likely rather weak in our work setting, we allow for reputation concerns in a reduced form way, to allow canonical models a chance to predict non-zero effort in periods without performance pay. We assume that in each period there is a constant probability that managers observe a worker, and choosing higher effort means that the manager will be more impressed, and more likely to re-hire the worker in the future, conditional on the worker being observed. The product of the probability of observation, and the benefit of impressing the manager, is denoted by $p(e_t)$, with $p(\cdot)$ increasing and concave. In all models we assume that the marginal utility of income is constant and unaffected by earnings, because the magnitude of earnings accumulated during a few hours of work is trivial relative to lifetime income. We normalize the marginal utility of income to 1.

Model (1): Canonical model:

In the simplest canonical model treatment group and control group workers maximize utility functions V_T and V_C:

$$V_T = z \cdot e_1 + w + p(e_1) - c(e_1) + w + p(e_2) - c(e_2) \tag{2}$$

$$V_C = w + p(e_1) - c(e_1) + w + p(e_2) - c(e_2). \tag{3}$$

Proposition 1: In a canonical model treatment group workers work harder than control when performance pay is present, but choose the same effort level once performance pay is removed.

It is straightforward to see that treatment workers have a higher optimal effort in period 1 than control workers, because performance pay increases the marginal benefit of effort. In period 2, however, the maximization problem for treatment and control workers, and thus optimal effort, is identical. Note that because the marginal utility of income is assumed to be unaffected by piece rate earnings, treatment workers have no greater taste for leisure in period 2 than control workers (no "standard inter-temporal substitution" effect).

[36] There is no need for a baseline period as any model predicts the same behavior for treatment and control in such a period. The representation of the worker's problem as involving one choice variable, effort (leisure), in a time-separable utility function that is linear in income, and convex in effort, is equivalent, along the optimal path, to a standard model of inter-temporal labor supply in which a worker has two choice variables, consumption and effort. Intuitively, the maximization problem in two variables can be reduced to a single variable problem by substitution of the first order condition for consumption; the convexity of effort costs in the resulting condition follows from concavity of utility in consumption (see, e.g., Browning et al, 1985; Fehr and Goette, 2007).

Model (2): Model with negative impact of performance pay on non-monetary motivation:

In this model we introduce non-monetary motives to work hard in a reduced form way, including an additional term in the utility function, θ, which increases the marginal utility of effort. This could arise for various reasons, for example because the task is enjoyable or because working hard satisfies a social norm. The crowding out hypothesis can be captured by assuming that non-monetary motivation is lower if the worker is experiencing, or has recently experienced, performance pay: $\frac{\partial \theta_t}{\partial z} < 0$. Crowding in is captured by assuming the opposite. Workers maximize the following:

$$V_T = \theta(z > 0)e_1 + ze_1 + w + p(e_1) - c(e_1) + \theta(z > 0)e_2 + w + p(e_2) - c(e_2) \tag{4}$$

$$V_C = \theta(0)e_1 + w + p(e_1) - c(e_1) + \theta(0)e_2 + w + p(e_2) - c(e_2) \tag{5}$$

Proposition 2: In the case of crowding out, with $\frac{\partial \theta_t}{\partial z} < 0$, and assuming $-\frac{\partial \theta_t}{\partial z} < 1$, treatment group workers exert more effort than control while performance pay is present, and less effort than control once performance pay is removed. In the case of crowding in, with $\frac{\partial \theta_t}{\partial z} < 0$, treatment workers exert more effort than control when performance pay is present, and more effort once performance pay is removed.

Intuitively, with crowding out, non-monetary motives are lower in period 1 if there is performance pay. This works against a positive effect of financial incentives, but if incentives are strong enough, output will still increase in period 1. A necessary condition for an increase in period 1 effort is that the reduction in period 1 non-monetary motivation, $-\frac{\partial \theta_t}{\partial z}$, is less than the marginal utility of income, which in this case is set equal to 1. In period 2, the reduction in non-monetary motivation caused by the previous experience of performance pay unambiguously lowers effort of treatment group workers relative to control group workers, because unlike in period 1 there are no offsetting financial incentives. In the case of crowding in, non-monetary benefits of effort increase in period 1, reinforcing financial incentives. Non-monetary marginal benefit is also higher for treatment workers than control in period 2, while financial motives are the same, so effort unambiguously increases.

Model (3): Model with fatigue spillovers:

Next, we modify the canonical model to allow for fatigue. In quite general terms, fatigue can be though of as a stock that increases the marginal cost of effort. The stock should be higher if the worker chose high effort in the previous period, and lower or zero if the worker rested instead. Specifically, we assume the same convex cost function as in the canonical model, except that the cost of effort in period 2 also depends on k, a fatigue stock: $c = c(e_2, k)$. We assume $\frac{\partial c(e_2,k)}{\partial e_2 \partial k} > 0$, so that the marginal cost of effort in period 2 is increasing in the fatigue stock.

We first consider a case with no rest break between period 1 and period 2, so that higher period 1 effort increases the fatigue stock for period 2: $\frac{\partial k(e_1)}{\partial e_1} > 0$. Treatment and control workers maximize the following:

$$V_T = ze_1 + w + p(e_1) - c(e_1) + w + p(e_2) - c(e_2, k(e_1)) \tag{6}$$

$$V_C = w + p(e_1) - c(e_1) + w + p(e_2) - c(e_2, k(e_1)). \qquad (7)$$

Proposition 3: In the case of fatigue spillovers and no rest break, if treatment group workers exert more effort than control while performance pay is present, they reduce effort relative to control once performance pay is removed.

Workers are assumed to be forward looking and take into account that period 1 effort makes it harder to exert effort in period 2. If financial incentives are strong enough, however, it will make sense for treatment workers to increase period 1 effort relative to control workers, even though this makes it harder to exert effort later on.[37] In this case it is clear that treatment workers have lower optimal effort than control workers in period 2, because their marginal cost of effort is higher in period 2.

If there is a sufficient rest break between period 1 and period 2, however, higher effort in period 1 for treatment workers need not imply lower effort than control in period 2. Sufficient means a rest break long enough to reduce the fatigue stock to zero by the beginning of period 2. In this case $k(e_1) = 0$ and $\frac{\partial k}{\partial e_1} = 0$ and we have the following proposition.

Proposition 4: With fatigue spillovers, but also a sufficient rest break following the performance pay episode, treatment group workers may exert more effort than control when performance pay is present, but the same as control after performance pay is removed.

With a sufficient rest break the cost of effort in period 2 is just $c(e_2)$, and the optimization problem facing the worker is the same as in the canonical model. Thus, the model predicts equal effort for treatment and control group workers in period 2. It is an empirical question whether the work task had significant fatigue spillovers and whether the rest break was long enough to eliminate any fatigue stock built up due to extra effort under performance pay. We investigate this question is various ways in the analysis.[38]

In summary, a canonical model would not predict the signature feature of the classic crowding out pattern. A model with intrinsic motivation, and crowding out, unambiguously predicts lower effort for treatment than control after performance pay is removed, and higher effort while performance pay is present if the marginal utility of income exceeds the marginal reduction in intrinsic motivation. With crowding in, the model predicts higher effort for treatment in both periods. Adding fatigue spillovers to the canonical model can generate the crowding out pattern, but if there is a sufficient rest break, predictions revert back to the canonical case.

[37] The marginal benefit of higher effort in period 1 needs to offset the marginal cost in period 1 as well as the increased marginal cost in period 2.

[38] The model also predicts a downward sloping effort profile for both control and treatment workers. As shown in Goette and Huffman (2006), however, fatigue effects do become more complex in a model with more than two periods. For example, forward looking workers might exhibit a u-shaped effort profile if there are three or more periods: at the beginning of the day there is no fatigue stock, so marginal cost is low, and the worker puts in some extra effort; during the middle of the work day the worker paces him or herself; before a rest break, or before the end of the workday, the worker might increase effort again, knowing that the future rest period wipes out the consequences in terms of accumulated fatigue stock. We abstract away from such effects in the analysis, focusing on more basic predictions of the fatigue model, e.g., that if treatment workers have higher effort in a given period, they should exert less effort in the next effort than control workers, all else equal.

A.0.2 Proofs

This portion of the appendix provides proofs of the propositions in the behavioral predictions section.

Proof of Proposition 1 First order conditions for a treatment group worker are given by:

$$\begin{aligned}\frac{\partial v}{\partial e_1} &= z + p'(e_1) - c'(e_1) = 0\\ \frac{\partial v}{\partial e_2} &= p'(e_2) - c'(e_2) = 0\end{aligned}$$

By inspection, optimal effort for treatment workers in period 1, e_1^{*T}, is increasing in the piece rate, given concavity of $p(\cdot)$ and convexity of $c(\cdot)$. Optimal effort in period 2, e_2^{*T}, however, is independent of the piece rate. First order conditions for a control group worker are given by:

$$\begin{aligned}\frac{\partial v}{\partial e_1} &= p_1' - c'(e_1) = 0\\ \frac{\partial v}{\partial e_2} &= p_2' - c'(e_2) = 0\end{aligned}$$

Optimal effort in period 1 for control workers, e_1^{*C} is obviously less than optimal effort for treatment workers, given that control workers have $z = 0$ in the first order condition for period 1. The first order condition for period 2 effort is identical for treatment and control so we have $e_2^{*C} = e_2^{*T}$ regardless of the level of z. Thus, $z > 0$ causes treatment group workers to work harder in period 1 than control group workers, but effort is the same for treatment and control in period 2.

Proof of Proposition 2:

First order conditions for a treatment group worker are given by:

$$\begin{aligned}\frac{\partial v}{\partial e_1} &= z + \theta_1 + p_1' - c_1' = 0\\ \frac{\partial v}{\partial e_2} &= \theta_2 + p_2' - c_2' = 0\end{aligned}$$

Totally differentiating with respect to z yields

$$\begin{bmatrix} c_1'' - p_1'' & 0 \\ 0 & c_2'' - p_2'' \end{bmatrix} \begin{bmatrix} \frac{\partial e_1}{\partial z} \\ \frac{\partial e_2}{\partial z} \end{bmatrix} = \begin{bmatrix} 1 + \theta_1' \\ \theta_2' \end{bmatrix}$$

Denote the first matrix by H. Note that, due to convexity of $c(\cdot)$ and concavity of $p(\cdot)$, $|H| = (c_1'' - p_1'')(c_2'' - p_2'') > 0$.

Applying Cramer's rule, the derivatives of the first and second period effort levels with respect to the piece rate are given by

$$\frac{\partial e_1}{\partial z} = \frac{\det \begin{bmatrix} 1 + \theta_1' & 0 \\ \theta_2' & c_2'' - p_2'' \end{bmatrix}}{|H|} \quad \text{and} \quad \frac{\partial e_2}{\partial z} = \frac{\det \begin{bmatrix} c_1'' - p_1'' & 1 + \theta_1'' \\ 0 & \theta_2'' \end{bmatrix}}{|H|}$$

In order to have $\frac{\partial e_1}{\partial z} > 0$ and $\frac{\partial e_2}{\partial z} < 0$, the signs of the determinants of the two numerator matrices must be positive and negative, respectively. Writing out the determinants, these conditions can be stated

$$(1+\theta_1')(c_2''-p_2'') > 0 \qquad \text{and} \qquad (c_1''-p_1'')\theta_2' < 0.$$

The first condition holds as long as $1 > -\theta_1'$. Recall that the marginal utility of income was normalized to 1. Thus, the first condition states that period 1 effort is increasing in the wage, as long as the marginal utility of an additional dollar is greater than the marginal reduction in intrinsic motivation from introducing $z > 0$. The second condition always holds, given our assumptions: introducing performance pay (for period 1) must reduce period 2 effort because it decreases intrinsic motivation in that period ($\frac{\partial \theta_2}{\partial z} < 0$), without generating any offsetting financial incentives in period 2. This proves the claim in Proposition 4.

Proof of Proposition 3:

First order conditions for a treatment group worker are given by:

$$\frac{\partial v}{\partial e_1} = z + p'(e_1) - c'(e_1) - c'(e_2, k(e_1))k'(e_1) = 0$$
$$\frac{\partial v}{\partial e_2} = p'(e_2) - c'(e_2, k(e_1)) = 0$$

Totally differentiating with respect to z yields:

$$\begin{bmatrix} c''(e_1) + c''(e_2,k(e_1))k'(e_1) - p''(e_1) & c''(e_2,k(e_1))k'(e_1) \\ c''(e_2,k(e_1))k'(e_1) & c''(e_2,k(e_2)) - p''(e_2) \end{bmatrix} \begin{bmatrix} \frac{\partial e_1}{\partial z} \\ \frac{\partial e_2}{\partial z} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Denote the first matrix, consisting of second derivatives, by H. Second order conditions for an (interior) maximum imply that the determinant of H must be positive. Applying Cramer's Rule, the derivatives of the first and second period effort levels with respect to z are given by

$$\frac{\partial e_1}{\partial z} = \frac{\det \begin{bmatrix} 1 & c''(e_2,k(e_1))k'(e_1) \\ 0 & c''(e_2,k(e_2)) - p''(e_2) \end{bmatrix}}{|H|} \qquad \text{and} \qquad \frac{\partial e_2}{\partial z} = \frac{\det \begin{bmatrix} c''(e_1) + c''(e_2,k(e_1))k'(e_1) - p''(e_1) & 1 \\ c''(e_2,k(e_1))k'(e_1) & 0 \end{bmatrix}}{|H|}$$

In order to have $\frac{\partial e_1}{\partial z} > 0$ and $\frac{\partial e_2}{\partial z} < 0$, the signs of the determinants of the two numerator matrices must be positive and negative, respectively. Writing out the determinants, these conditions can be stated

$$c''(e_2,k(e_2)) - p''(e_2) > 0 \qquad \text{and} \qquad -c''(e_2,k(e_1))k'(e_1) < 0.$$

Both of these hold unambiguously given assumptions about convexity of $c(\cdot)$, concavity of $p(\cdot)$, and $k'(e_1) > 0$. This proves the claim in Proposition 2.

Proof of Proposition 4:

With a rest break between period 1 and period 2 sufficient to have a fatigue stock of zero regardless of period 1 effort, we have $k'(e_1) = 0$. As shown in the proof for Proposition 2 this implies $\frac{\partial e_2}{\partial z} = 0$, so period 2 effort of treatment workers is unaffected by having performance pay in period 1. Given $c(e_2, 0) = c(e_2)$, treatment and control workers have the same first order conditions and optimal effort levels for period 2 effort. This proves the claim in Proposition 3.