For love or money? Comparing the effects of non-pecuniary and pecuniary incentive schemes in the workplace[1]

Omar Al-Ubaydli, Steffen Andersen, Uri Gneezy and John A. List[2,3]

October 2008

Abstract

Constructing compensation schemes for effort in multi-dimensional tasks is complex, particularly when some dimensions are not easily observable. When pecuniary schemes contractually reward workers for their observable effort, the unrewarded dimensions might be neglected relative to the optimal solution. An alternative solution is to use non-pecuniary schemes, such as gift exchange in which the employer pays the worker above market clearing wages and the workers with social preferences reciprocate. We compare pecuniary and non-pecuniary incentive schemes side-by-side in a natural field experiment where we have accurate measures of all dimensions of effort, unbeknownst to workers. We find that workers' responses to both types of incentive schemes are largely consonant with theory. Interestingly, we detect an asymmetrical response to non-pecuniary gifts: workers focus on observable effort at the expense of diminished unobservable effort when reciprocating positive gifts, yet they decrease all dimensions of effort in response to negative gifts.

JEL codes: D63, D82, J3
Keywords: Gift exchange, piece-rate, incentives

[1] We wish to thank Min Sok Lee for his help in running these experiments.
[2] Al-Ubaydli: Department of Economics and Mercatus Center, George Mason University. Andersen: Centre for Economic and Business Research, Copenhagen Business School. Gneezy: Rady School of Management, University of California at San Diego. List: Department of Economics, University of Chicago and NBER.
[3] Corresponding author: Omar Al-Ubaydli, George Mason University Department of Economics, 4400 University Drive, MSN 3G4, Fairfax, VA 22030, USA. Tel: +1-703-993-4538, email: omar@omar.ec

> **“**A Father, being on the point of death, wished to be sure that his sons would give the same attention to his farm as he himself had given it. He called them to his bedside and said, "My sons, there is a great treasure hid in one of my vineyards." The sons, after his death, took their spades and mattocks and carefully dug over every portion of their land. They found no treasure, but the vines repaid their labor by an extraordinary and superabundant crop.” Aesop’s Fables.
>
> “Do not withhold good from those to whom it is due when it is in your power to do it.” Proverbs 3:27.

I. Introduction

Designing optimal incentive schemes is perhaps one of mankind’s oldest activities. From the Dead Sea Scrolls to scribes on tombs of ancient kings, rudimentary and clever incentive structures to motivate a particular course of action have been extolled. For their part, economists have produced a rich assortment of models that lend insights into the various factors that are likely to influence equilibrium market behavior. A particularly important and relatively complex problem arises when output has multiple dimensions that vary in their quantifiability. For example, school teachers are responsible for improving a wide range of students’ academic and social skills, most of which are difficult to objectively measure. The folk theorem formalizes how reputational concerns can circumvent such multitasking problems. Employer-worker interactions, however, often lack the necessary intensity or horizon forcing employers to devise alternative incentive schemes.

The seminal theories addressing optimal incentive schemes in multitasking were put forward by Holmstrom and Milgrom (1991) and Baker (1992). They start by explaining why simply rewarding workers on the easily measurable components may fail: workers may systematically neglect the unrewarded components of output to the detriment of the employer. This intuition is subsequently used to explain the prevalence of flat-wage structured incentives in occupations where piece-rates could be easily implemented.

Akerlof (1982) proposed gift exchange as an alternative solution to the multitasking problem. The idea is that even when a worker is paid a flat wage, the level of pay influences effort, even in situations void of reputational considerations. In contrast to the traditional assumptions in economic models, workers are not mute to reciprocal motivations in the gift exchange model. Once receiving a gift in the form of higher than market clearing wages, workers reciprocate with

higher effort levels (positive reciprocity); when receiving less than they believe they deserve, workers punish the firm via sabotage or other more subtle measures (negative reciprocity). The reliance of such schemes on non-egotistical preferences suggests that they can be thought of as non-pecuniary, despite having a pecuniary component in the literal sense.[4]

In this paper, we compare pecuniary and non-pecuniary schemes as solutions to the multitasking problem using a natural field experiment. The prospect to design and test the efficacy of various incentive schemes in an actual work environment presented itself when we agreed to help a research organization with a capital campaign in which 5,500 potential donors were to receive direct mail solicitations.[5] The task to be completed – preparing materials for distribution by the charity – permits us to measure output along two crucial dimensions (quantity and quality) despite workers thinking that we could only measure quantity. This is because they were not aware that part of the project's motivation was research into labor productivity. Therefore, we are able to provide a two-dimensional test of our incentive schemes, capturing the essential elements of the relevant multitasking theory.

Our study also has several other innovations. First, by implementing positive and negative shocks to worker compensation, we are permitted a novel comparison of positive and negative reciprocity. Existing studies usually restrict attention to one and do not consider multidimensional output. Second, we use a novel subject pool: temporary workers.[6] Such workers are likely to have reputational concerns that represent an interesting middling ground between subjects in one-shot experiments and full-time employees in firms.

We report several insights. First, in this particular work environment, pecuniary incentives have the predicted effect when compared to flat wage schemes: quantity levels increase at the expense

[4] This conjecture is typically termed the "fair wage-effort" hypothesis. An alternative underlying mechanism that can create similar data patterns is denoted the "efficiency wage theory," which surmises that wages above market-clearing levels occur because these wage profiles induce workers to be motivated in an effort to avoid being fired, which economizes on firm-level monitoring (see, e.g., Katz, 1986).

[5] The Natural Hazard Mitigation Research Center was authorized to begin operations in the fall of 2004 by the North Carolina state government. The Hazard Center was founded in response to the widespread devastation in Eastern North Carolina caused by hurricanes Dennis and Floyd, and designed to provide support and coordination for research on natural hazard risks. For more information on the Hazard Mitigation Research Center see www.artsci.ecu.edu/cas/auxiliary/hazardcenter/home.htm.

[6] Professional temporary workers are widespread. In 2005 alone, US staffing companies employed an average of 2.9 million temporary and contract workers. Further, on any given day the staffing industry employs more than 2% of the US work force (American Staffing, 2006).

of lower quality. Second, non-pecuniary incentives also seem to have an effect on both output dimensions. Interestingly, workers exhibit an asymmetry in the nature of their reciprocity. When receiving a positive gift, their response focuses on the observable component of output: quantity increases yet quality *decreases*. In contrast, when receiving a negative gift, they reciprocate negatively in *both* output dimensions. Third, the estimated treatment effects were in some cases substantially smaller than we would have expected from parallel laboratory experiments. Amongst other reasons, we believe this to be the result of the reputation effect of hiring workers through a temporary employment agency: the quality and frequency of future job-offers from the agency to its workers are a function of past performance, e.g., a worker who fails to show-up to a job without prior warning or who performs poorly is struck from the agency's database.

The remainder of our study proceeds as follows. Section II is the background and existing literature. Section III is the experimental design. Section IV is the results. Section V concludes.

II. Preliminaries

A. Theory

The theory that follows is based on Holmstrom and Milgrom (1991) and Baker (1992). We relegate the derivation to the appendix. The model's key features include:

- Effort t is two-dimensional but only t_1 is observable to the principal.
- Benefit to principal of effort is $B(t)$. Cost to agent is $C(t)$. $C(t)$ has an interior minimum for some finite strictly positive vector $\bar{t}$ (due to, e.g., boredom; see Holmstrom and Milgrom (1991)).[7]
- Compensation scheme: $w = \alpha + \beta t_1$.
- Preferences:
    - Principal $u_p = B(t) - w$.
    - Agent $u_a = w - C(t) + \rho k u_p$. k is how kind the agent thinks that the principal is being, with $k = 0$ denoting neutrality.[8] $\rho \geq 0$ is the agent's reciprocity parameter; $\rho = 0$ implies a neoclassical agent, $\rho > 0$ implies a reciprocal agent.

[7] $B(t)$ is strictly increasing and strictly concave. It is continuously differentiable with $B_i \to \infty$ as $t_i \to 0$ and $B_i \to 0$ as $t_i \to \infty$. $C(t)$ is strictly convex. It is continuously differentiable with $C_i \to -\infty$ as $t_i \to 0$ and $C_i \to \infty$ as $t_i \to \infty$.

- Under regularity conditions, the solution to the worker's problem is $t^*(\beta, \rho k)$.

Prediction 1 [Multi-tasking problem]: A neoclassical agent responds to a flat wage with selfish effort: $t^*(0,0) = \bar{t}$.

Typically, $\bar{t}$ is sub-optimal from the principal's perspective.

Prediction 2 [Pecuniary scheme]: If $C_{12} > 0$, a neoclassical agent responds to a piece-rate by increasing observable effort at the expense of unobservable effort: $t_1^*(\beta, 0)|_{\beta>0} > \bar{t}_1$ and $t_2^*(\beta, 0)|_{\beta>0} < \bar{t}_2$.

$C_{12} > 0$ means that the marginal cost of one effort is increasing in the other effort. This is a natural assumption when there is substitutability in the effort in the task.[9] When β increases, it is obviously in the agent's interest to increase t_1 since he is directly rewarded for doing so. Since $C_{12} > 0$, by decreasing t_2, he lowers the marginal cost of increasing t_1 so that he can afford to increase it even further. Hence, if t_2 is very important to the employer, then it is optimal to use a flat wage profile when faced with neoclassical workers.

Prediction 3 [Pecuniary scheme]: Suppose that agents differ in their ability A and that the cost of effort is $\frac{C(t)}{A}$. Then for neoclassical agents, the variance of both efforts increases: $\frac{\partial Var(t_i^*)}{\partial \beta} > 0$ (Lazear (2000)).

The higher the performance related coefficient of compensation, the larger the incentive for those with high ability to implement that higher ability.

The basis of Akerlof's (1982) gift exchange model is the idea that, even in one-shot environments, if the agent thinks that the principal is giving him a better-than-fair deal $(k > 0)$, he is willing to expend resources returning the favor. Likewise, the agent is also willing to expend resources to spite the principal if he believes that the principal is providing an "unfair" wage profile $(k < 0)$.

---

[8] We discuss kindness and reciprocity more below. See Dufwenberg and Kirchsteiger (2004) and Fehr and Fischbacher (2006).

[9] For example, suppose that the agent is building cars, with t_1 representing the number of cars produced and t_2 representing the number of times he checks for defects. The more time he checks for defects, the higher the cost of exerting the effort to produce a certain number of cars.

Assumption: The market compensation corresponds to neutral kindness ($k = 0$), and paying above the market compensation increases kindness: $\frac{\partial k}{\partial \alpha} > 0$.

Prediction 4 [Non-pecuniary scheme]: If $B_{12} - C_{12}$ is sufficiently small, a reciprocal agent responds to increasing compensation by increasing both observable and unobservable effort: $\frac{dt_i^*(\beta,\rho k(\alpha))}{d\alpha}\Big|_{\rho>0} > 0$.

Despite the financial component to such incentive schemes, we can view them as being non-pecuniary because they *rely on social preferences*.[10] Whenever k increases, the agent's preferences are more aligned with the principal's. The larger B_{12}, the greater the complementarity in effort levels, and so the greater the gain to the principal of increasing each effort. The smaller C_{12}, the smaller the adverse effect to the agent of increasing effort on the marginal cost of the other. Whether or not it is profitable to pay above market wages depends on the size of ρ.

Prediction 5 [Non-pecuniary scheme]: If $C_{12} > 0$ and the agent reciprocates in observable effort only ($u_a = w - C(t) + \rho k t_1$), then a reciprocal agent responds to increasing compensation by increasing observable effort at the expense of unobservable effort: $\frac{dt_1^*(\beta,\rho k(\alpha))}{d\alpha}\Big|_{\rho>0} > 0$ and $\frac{dt_2^*(\beta,\rho k(\alpha))}{d\alpha}\Big|_{\rho>0} < 0$.

This is in anticipation of some our empirical results. When the agent reciprocates in observable effort only, increasing k is equivalent to increasing β. Therefore the efficacy of gifts depends on both ρ and on the extent to which the agent reciprocates in the u_p vis-à-vis t_1.

As a final note, the above theories are formulated for one-shot settings. As the intensity and horizon of interactions between the principal and agent increase, the Folk theorem reminds us that opportunistic behavior can be eliminated by the threat of mutual enforcement. As such, under some circumstances, increasing reputational considerations diminishes the need for the pecuniary and non-pecuniary incentive schemes described above.

[10] Recall that this is a one-shot setup and so the mechanism is unrelated to reputational effects.

B. Empirical evidence

Several studies have demonstrated increased observable effort in response to the introduction of a piece-rate incentive scheme.[11] There is also a large body of laboratory experimental evidence consonant with gift exchange, though some studies cast doubt on its robustness.[12] Yet survey evidence also supports the notion that wage and effort combinations can be above the neoclassical equilibrium.[13] Evidence from field experiments also provides mixed support of a long-term gift effect.[14]

We extend the literature by providing a two-dimensional test of the prominent incentive schemes, capturing the essential elements of multitasking theory. Existing studies that explore quality typically only use data that is available to the manager and can in principle be contracted upon (which it actually is in the case of Paarsch and Shearer (2000) and Lazear (2000)). In our experiment, the subjects are unaware that they are participating in a labor productivity study, knowing only that they have been hired by a non-profit organization to pack envelopes. As such, they do not anticipate our extremely costly checking of every envelope, detailed below. Moreover, as a consequence of this painstaking procedure, we are able to collect data on a very rich array of quality measures.

Our approach also extends the literature by using temporary workers rather than laboratory subjects in a one-shot setting or field subjects who are full-time employees. We expect temporary workers to have intermediate reputational concerns. This follows from the fact that the frequency and quality of jobs that are offered to a temporary worker – unlike one-shot subjects – is typically a function of past performance.[15] For example, should a worker fail to show up to a job without prior warning, he or she is struck from the agency's database. Similarly, if the agency receives a bad report about a worker from a hiring firm, the worker's future prospects suffer. However this reputation effect likely has less power than the reputational concerns of

[11] See Lazear (2000), Johnson et al. (2006), Paarsch and Shearer (2000), Shearer (2004).
[12] See Fehr et al. (1993, 1997) and Charness (2005); Engelmann and Ortmann (2002), Charness et al. (2004).
[13] See Agell and Lundborg (1995). Making use of naturally-occurring data, Krueger and Mas (2004) and Mas (2006) find evidence consistent with negative reciprocity.
[14] See Pritchard et al. (1972), Gneezy and List (2006), Kube et al. (2006), Hennig-Schmidt et al. (2006); for naturally occurring data, see Chen (2005) and Lee and Rupp (2006).
[15] It is common to use temporary workers for a task such as envelope packing.

workers in full-time jobs, and so presents an interesting "middling" testing environment for the various incentive schemes proposed.

Finally, to the best of our knowledge, we are the first study to compare pecuniary and non-pecuniary schemes in a consistent environment, as well as positive and negative reciprocity in a consistent field environment.

III. Experimental design

As aforementioned, we agreed to work jointly with the Center for Natural Hazards Mitigation Research to help them raise capital. The task was the preparation of solicitation letters, a multitasking problem. The goal of this experiment is to compare pecuniary and non-pecuniary incentive schemes to the baseline of a market flat hourly wage in a natural environment. We believe that exploring behavior in a controlled, real-world setting which theory attempts to explain is an important next step in the literature. Our employees complete real effort tasks, they are not told explicitly the profit distributions, and they are likely to have previous experience working in similar environments. Beyond learning about non-pecuniary and pecuniary incentive schemes in such a setting, we view this approach as providing insights into the underpinnings of a worker pool that has not been studied in a controlled experiment. According to the American Staffing Association's quarterly employment and sales survey, temporary workers are quite broad-based.[16]

A. Treatments

Table 1 summarizes our experimental design. We had two baseline treatments where workers were hired at, and paid, a flat hourly wage of $8: *1-day-8* (lasting one day) and *2-day-8* (lasting two days).[17] In the one-day treatment, employees were contracted to work from 9am-3pm, which included 30 minutes of orientation/late arrivals and a 30-minute paid lunch-break. Thus, effectively, each worker is on the job for five hours. This was identical for the first day of the

[16] Roughly 2.9 million workers are employed via temporary agencies, and sales for staffing in 2005 totaled $69.5 billion. Further, approximately one in eleven non-farm workers had a job with a staffing company at some point in 2005.

[17] We learned that $8/hr is approximately the local market clearing wage rate for a typical envelope-packing task, thus we made use of the market clearing wage in our baseline treatment.

two-day treatment. On the second day, employees worked for two consecutive hours.[18] In the two-day treatments, before work commenced on day one, workers were informed that they would be invited back on the second day should their performance on the first day be deemed satisfactory.[19]

Table 1: Experimental design summary

Treatment	Date	Number of Workers	Hiring compensation	Actual compensation
1-day-8	6/7/2006	26	$8/hr	$8/hr
1-day-16	6/13/2006	27	$16/hr	$16/hr
2-day-8	6/15/06-6/16/06	24	$8/hr	$8/hr
2-day-positive	6/19/06-6/20/06	29	$8-16/hr	$16/hr
2-day-negative	6/26/06-6/27/06	30	$8-16/hr	$8/hr
1-day-piece	6/22/2006	29	$6.50/hr + $0.15/envelope	$6.50/hr + $0.15/envelope
2-day-piece	6/29/06-6/30/06	26	$6.50/hr + $0.15/envelope	$6.50/hr + $0.15/envelope

Treatment designation denotes days of work and payment scheme. For example, "1-day-8" denotes one day of work at a wage rate of $8 per hour.

We implemented one- and two-day treatments for several reasons. First, it offered us an opportunity to test whether additional reputational concerns, beyond the reputational effects associated with the temporary agency effect, are important. Second, Gneezy and List (2006) found that in two different tasks, the effects of non-pecuniary schemes were quite strong at first before tapering off to insignificance within a few hours. Thus a two-day treatment acted as a robustness check. Finally, the two-day treatments offered us an opportunity to implement unanticipated shocks to the compensation schemes by virtue of extending the number of working hours (see section IV).

[18] In fact, we actually employed the workers for four hours on the second day. However we shocked their labor contracts during the last two hours. We report the results of these shocks in section IV below.
[19] No worker was terminated after the first day.

To test pecuniary schemes, we used two separate piece-rate treatments. In Table 1 these are denoted as *1-day-piece* and *2-day-piece*. During hiring in both treatments, workers were told that they would be paid a minimum wage ($6.50/hr) plus a per-envelope-bonus. When workers arrived, and after the task was explained, they were told that the per-envelope-bonus was $0.15.[20] This figure was chosen because in the *1-day-8* and *2-day-8* treatments, average productivity was roughly ten envelopes per person; therefore, the piece-rate loosely corresponded to a wage rate of $8 per hour.[21] Similar to the two day treatments described above, workers in the *2-day-piece* treatment were informed that they would be invited back on the second day should their performance on the first day be deemed satisfactory.

Testing non-pecuniary schemes was slightly more problematic since there is no clear recipe on manipulating the kindness term in recent intentions-based reciprocity models of, for example, Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006). One way would be to simply advertise and pay a wage that is above market-clearing. This was what we did in *1-day-16*, where workers were hired at, and paid, an hourly wage of $16 for a one-day job. An alternative was to use the surprise method in Gneezy and List (2006). In *2-day-positive*, workers were informed during the hiring phase that they would be paid somewhere in the region of $8-$16/hr. Further, the prospective workers were informed that they would learn their actual pay when they arrived on the job. When they arrived, after the task was explained to them, the workers were told that their wage was $16/hr.

If we find that there is no selection effect in hiring or additional reputation effect associated with two-day jobs, then this treatment also provides an interesting comparison with *1-day-16*. In this case, any differences should be driven by the surprise nature of the gift, permitting us to parse the importance of the pure wage effect and the surprise effect.

We also wanted to explore negative reciprocity – reciprocity is a double-edged sword. In *2-day-negative*, during the hiring phase again workers were told that they would be paid somewhere in

[20] We can envision that the piece-rate schemes might lack power since the wage is composed of a combination fixed and variable wage rather than the higher powered purely piece rate method. We did attempt to use a pure piece-rate, but we were informed by the temporary employment agency that minimum-wage laws in Illinois forbade such an approach.

[21] Ten envelopes per hour may appear to be a low figure, yet the task contained a time-consuming administrative component after the packing of each envelope. See below for further details.

the region $8-$16 per hour, and that they would be informed of their actual pay when they arrived at the job. When they arrived, after the task was explained to them, the workers were told that their wage was $8/hr.22

To complement these treatments, we implemented unanticipated shocks to the labor contracts near the end of the work contract. For example, in 2-day-8, after two hours of work on the second day (i.e., after the seventh hour), we informed the employees that in appreciation of their work, we would pay them for the two remaining hours but that they would only work for one more hour (i.e., they would receive $16/hr for the last hour).23 This is a gift in the spirit of the positive reciprocity treatment. Alternatively, in 2-day-positive, after three hours of work on the second day, we cut the workers' wages from $16/hr to $8/hr for the last hour of work. This is a shock in the spirit of the negative reciprocity treatment.

Finally, in the 2-day-piece treatment we increased the wage rate for the last two hours from $6.50/hr + $0.15/envelope to $6.50/hr + $1/envelope (the new compensation only applied to the envelopes produced in the last two hours, and this was made clear to the workers). This was to intensify the piece-rate incentive, though it should be noted that since identical effort now yields a much higher income, it is impossible to determine how much of any change in behavior is driven by the shock being perceived as a gift.

B. Recruitment

In the Winter/Spring of 2006, we made inquiries to numerous employment agencies to hire workers to pack envelopes for our non-profit organization. We personally interviewed several agencies and selected the firm that we regarded as the best balance between the price charged and the flexibility required to run the natural field experiment.[24] In the end, we contracted with a local temporary-employment agency that was broad based and had years of experience in this area of work – FurstStaffing Services.

[22] While such a treatment surely will not invoke the type of negative reaction that a "promise of $16/hr and when they arrive tell them that they are only earning $8/hr" would yield, as in the *2-day-positive* treatment we followed standard protocol within the experimental community of not deceiving the subjects. Our results should therefore be interpreted accordingly. Moreover, we did not run a non-surprise negative reciprocity treatment (similar to *1-day-16*) since the market wage was so close to the minimum wage – a significant treatment effect did not seem likely.
[23] See the appendix for the exact wording.
[24] Without letting the agencies know about our plans, it was clear through these initial interviews that many of the agencies were not inclined to provide the necessary control to complete a controlled study.

We worked closely with the temporary work agency to recruit workers for our task in a natural manner. This was done by placing advertisements on various employment websites. The advertisements simply requested envelope packers for a non-profit organization at the University of Chicago, and requested that potential workers contact the employment agency if they had interest. Importantly, we followed standard procedure by not mentioning compensation or the exact date(s) of employment on the advertisements at this stage.

Recruitment generally followed four steps. First, when potential workers contacted the employment agency expressing interest in the job, the agency followed its usual protocol of inviting applicants in for a brief interview, which included completion of paperwork that provided a demographic profile. Second, workers were informed that the employment agency would contact them soon with further details about the job, including compensation and the date(s) at which employment was available.

These first two steps provided us with a large pool of potential workers. The (mutually exclusive) dates for each treatment were decided in advance and in step three, potential workers were randomly allocated to one of our seven treatments. In the fourth step, staff of the employment agency personally called each potential worker and inquired about their availability on their assigned day (corresponding to their allocated treatment). If the prospective employee replied in the affirmative, they were informed of the compensation package, and asked to confirm working for the specified compensation on the chosen date. If the prospective worker declined after learning about the wage, they were eliminated from the sample.[25]

If the prospective employee was not available to work on a specific date, while still being unaware of the compensation scheme, they were offered another randomly selected date. Importantly, no individual was offered more than one wage or made aware of the existence of a wage that differed from the one that they accepted or rejected.[26] The target recruitment for each treatment was between 25-30 workers; this necessitated hiring and confirming with 35-40 workers given an expected show-up rate of approximately 70%. One might conjecture that the higher wage rates would induce superior show-up rates, a question we explore more fully below.

[25] This provided us with a unique chance to explore aspects of sample selection into the various wage schemes. We discuss this issue further below.

[26] Working closely with us, the employment agency never made potential workers aware of the existence of multiple wage rates, attenuating any chance for cross-contamination.

C. The task

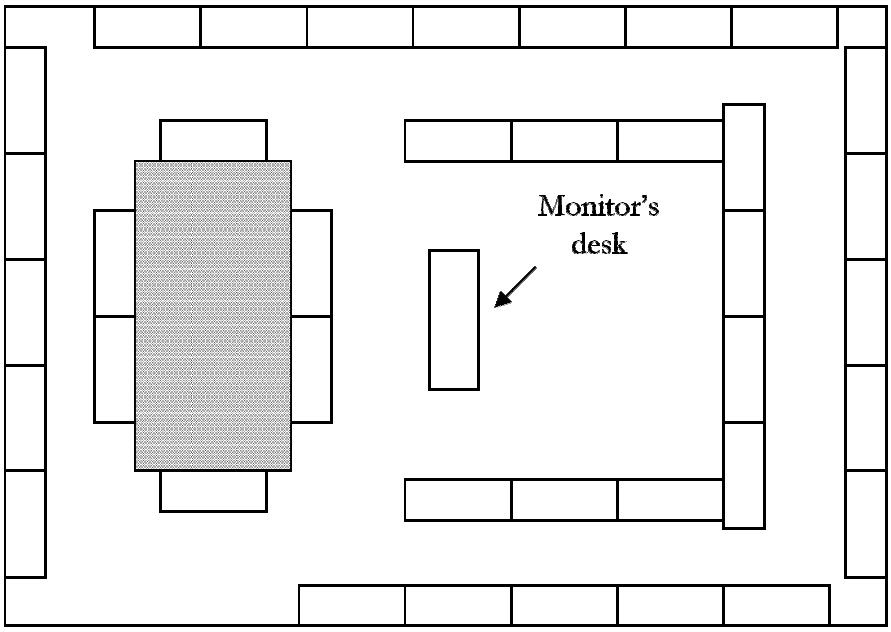
Upon arriving on site, each worker was greeted, signed-in, and seated at an individual six-foot desk. All workers in a treatment were in the same room and the desks were linked and outward facing (see Figure 1 for details concerning the seating arrangement).[27] The monitors' desk was placed in the center of the room, and this is also where supplies for the task were stored. Upon arrival of all workers, the monitors asked workers to be seated in front of the monitor's desk to receive instructions. The monitor proceeded with a 15-minute orientation.

Workers were informed that the task was part of a charitable fundraising drive. In this drive, several thousand letters needed to be prepared and mailed. The monitor informed workers that there was a stack of letters placed on their desk (an equal number on everyone's desk), as well as other support materials. The worker was to: (1) match the top letter from the stack with the matching address label, (2) read the letter to identify the correct accompanying materials, and (3) insert the necessary materials into the envelope. It was stressed that great importance lie in correctly producing and packing the envelopes. Workers were informed that they should then place completed envelopes to the side on their desk. We further informed workers that monitors would collect the completed envelopes during the day (in practice we collected them on the top of each hour).

In the final preparatory step workers completed important paperwork that accompanied each letter. This time-consuming task had the worker verify each letter that was sent to ensure that we did not send multiple letters to each household. Importantly, this task was simple to complete accurately, only requiring that the worker could read and write.[28] Yet, we did not collect individual paperwork until the close of the work day, sending a signal to the worker that this was a component of effort that would not be monitored as closely as the count of letters. The monitor closed the introduction by asking for questions. After all questions were answered, work commenced.

[27] There was no evidence of any seating effects of productivity (e.g., neighbour effects or position effects).
[28] Literacy was a requirement at the recruitment process that was covered by the employment agency.

Figure 1: Room layout and seating arrangement



The monitor's desk is where the supplies were stored too. Desks were approximately 6 feet wide and were outward facing.

In sum, workers were recruited to a natural work environment to complete a task for compensation. The task and wage profiles were similar to many other jobs they had accepted previously from employment agencies. The task involved a range of activities, each of which might include defects, or errors. This permits us to observe both the quantity and quality of the workers' output. In total, there are 12 potential errors, but the effect of each on output (the final output of an envelope is a successful contribution by the recipient of the specific letter) is uncertain.[29] We classified the errors into three broad categories reflecting the error's importance based on introspection.

- Critical errors: all but guaranteed failure of the letter to generate contributions. These included incorrectly matching the address label and letter (e.g., the household of Mr. J. Smith receives a letter addressed to the household of Mr. R. Jones) or forgetting to include either the address label or the letter.

[29] The letters were posted in August 2006. Below we discuss results from the mailer.

- Non-critical errors: represent a portmanteau for remaining errors in matching the materials, e.g., failing to include a complimentary bookmark when the letter states that the recipient should get one, or omitting a copy of the newsletter. This did not include making the mistake of putting extra materials into an envelope.

- Recording errors: reflect the administrative task that was completed directly after the envelope-packing task. As mentioned above, this was a time-consuming activity that required the workers to look up addresses in a long list. Errors of this class involve failure to complete it correctly or, as was often the case, especially in the piece rate, failure to attempt to complete the task. We viewed this error type as one that is not necessarily cognitive, rather it represented a simple task that would be time-consuming, and therefore yield fewer letters per hour.

Before proceeding to the empirical analysis, we should note a few important design features. First, workers were not informed that they were taking part in an experiment before, during, or after the exercise. Second, we were careful to remind workers that this was a one-time work opportunity. In cases where we employed workers for two days, they were informed that if their work was deemed acceptable they would be invited back for a second work day. Yet, as mentioned above, there is already in place a reputation effect from hiring via an employment agency. This may well impair our efforts to engineer a pure gift-exchange situation, a goal more convincingly attained by the large class of laboratory and field experiments noted in section II. One benefit from our setting is that the work environment in and of itself is much more natural, since purely one-shot employment is the exception rather than the norm in the field.

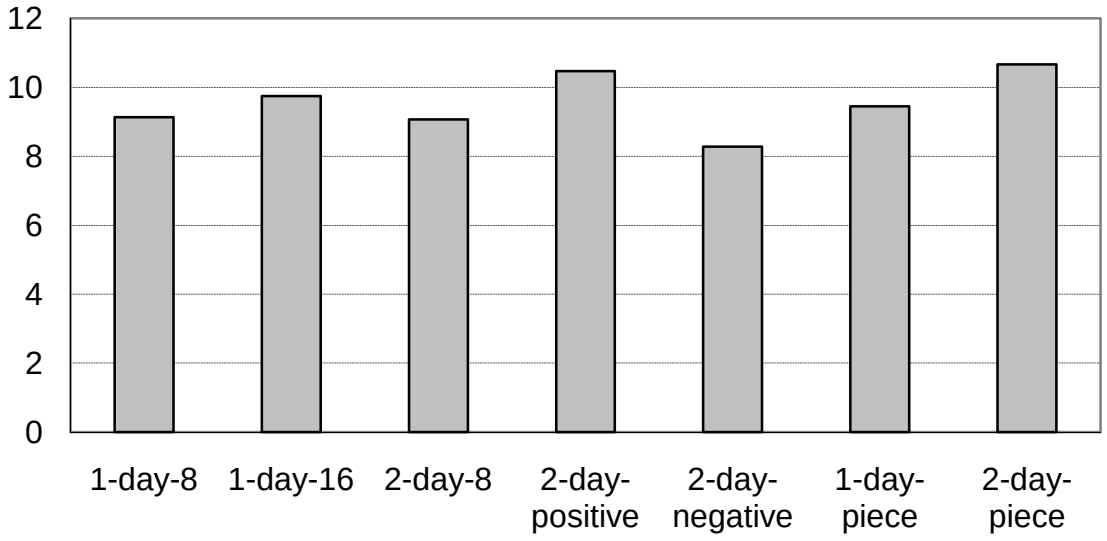
IV. Empirical results

For the remainder of this section, we divide our analysis into ‘results’ and ‘pararesults’. Results are conclusions that relate to the theoretical predictions from section II, while pararesults are ancillary conclusions. For brevity, we omit the explicit statistical inference underlying the pararesults. And, following convention we rely on statistical tests using two-sided alternative hypotheses even though our theory clearly provides one-sided predictions. This approach understates the significance of our results, but we leave it to the reader to make appropriate transformations of relevant p-values.

A. Main results

Figure 2 and Table 2 provide a summary of the raw data, which includes mean and standard deviation measures of our dependent variables and independent variables. Table 2 can be read as follows: in *1-day-8*, on average 9.131 letters were completed per worker per hour, with associated errors of 0.005 (critical), 0.235 (non-critical), and 0.304 (recording) per envelope. Further, roughly 77% of workers in this treatment were female, 81% were black, the average age was 35.5 years old, 61.5% of the sample received a high school diploma but did not receive education beyond high school, and 15.4% obtained a bachelor's degree.

Figure 2: Envelopes packed by treatment



Pararesult 1: Randomization across treatments was successful.

This is established either by using five one-way univariate ANOVAs (one for each demographic variable) or a unique one-way multivariate ANOVA with the five-dimensional demographic vector being the dependent variable. Both are statistically insignificant. For completeness we present results from both unconditional and conditional tests below.[30]

[30] The statistical tests and regressions presented below assume a treatment effect that does not differ with respect to socio-economic characteristics (such as race or gender). We examined the possibility of heterogeneous treatment effects, but we found only sporadic cases of statistical significance. Moreover, deviations from a homogenous treatment effect model did not exhibit any discernible pattern.

Table 2: Sample statistics by treatment

Variable	1-day-8	1-day-16	2-day-8	2-day-positive	2-day-negative	1-day-piece	2-day-piece
Panel A. Dependent variables							
Letters/hour	9.131	9.748	9.071	10.466	8.279	9.447	10.664
	(2.132)	(2.992)	(2.287)	(3.210)	(2.567)	(2.898)	(3.818)
Critical errors/envelope	0.005	0.043	0.003	0.001	0.004	0.012	0.014
	(0.011)	(0.192)	(0.008)	(0.006)	(0.013)	(0.035)	(0.022)
Non-critical errors/envelope	0.235	0.314	0.243	0.189	0.175	0.207	0.252
	(0.268)	(0.314)	(0.199)	(0.188)	(0.148)	(0.191)	(0.276)
Recording errors/envelope	0.304	0.259	0.095	0.200	0.210	0.280	0.196
	(0.426)	(0.357)	(0.214)	(0.341)	(0.342)	(0.403)	(0.278)
Panel B. Explanatory variables							
Female (dummy)	0.769	0.667	0.708	0.690	0.567	0.862	0.731
	(0.430)	(0.480)	(0.464)	(0.471)	(0.504)	(0.351)	(0.452)
Age in years	35.462	31.111	36.125	34.034	38.933	34.966	32.769
	(10.603)	(11.294)	(11.588)	(10.972)	(14.377)	(11.309)	(12.278)
Black (dummy)	0.808	0.778	0.875	0.793	0.800	0.828	0.846
	(0.402)	(0.424)	(0.338)	(0.412)	(0.407)	(0.384)	(0.368)
High school maximum (dummy)	0.615	0.778	0.625	0.724	0.533	0.621	0.615
	(0.496)	(0.424)	(0.495)	(0.455)	(0.507)	(0.493)	(0.496)
Bachelor's degree maximum (dummy)	0.154	0.074	0.208	0.207	0.267	0.276	0.231
	(0.368)	(0.267)	(0.415)	(0.412)	(0.450)	(0.455)	(0.430)

Figures displayed are means with standard deviations in parentheses. Dependent variables denote the four dimensions of productivity which we are seeking to explain in the statistical analysis. Letters / hour denotes the average letters packed by each worker, averaging across the workers and the hours worked. The three errors are averages across workers and envelopes packed. Critical errors are errors that render the output useless. Non-critical errors limit the usefulness of the output but do not necessarily eliminate it. Recording errors denote errors in an ancillary administrative task that has no direct effect on the usefulness of packed envelopes. Explanatory variables denote the conditioning variables in the statistical analysis. 'High school maximum' denotes workers whose highest academic qualification is a high school degree. See Table 1 for treatment definitions.

A first important question is whether the additional reputational effects within our flat wage treatments led to behavioral responses. Recall that all of our workers were informed early on that this is one time work; yet, even in the one day treatments effort is potentially rewarded by the employment agency because positive feedback likely leads to future employment opportunities. We augmented these reputation concerns in our two-day treatments by informing workers that they would be invited back on the second day "should their performance on the first day be deemed satisfactory." We purposely made the statement vague to explore how the additional concern for reputation influenced outcomes.

Pararesult 2: The distribution of output, critical errors and non-critical errors does not differ between *1-day-8* and *2-day-8*, but the distribution of recording errors does, being much lower in *2-day-8* than *1-day-8* (0.01 vs. 0.30).

Pararesult 3: The distribution of critical errors, non-critical errors and recording errors does not differ between *1-day-piece* and *2-day-piece*, but the distribution of output does, being higher in *2-day-piece* than *1-day-piece* (10.6 vs. 9.4).

These results are based on conditional and unconditional statistical test (parametric and non-parametric). In the remaining analysis, therefore, we pool the data from the *1-day-8* and *2-day-8*, except in the recording error case, where we present the non-pooled results; and we pool the data from *1-day-piece* and *2-day-piece*, except in the output case, where we present the non-pooled results. Note that in the cases where we do not pool, the difference is in the direction that theory would predict, i.e., better performance in the 2-day treatment.

An important question revolves around whether selection into the various treatments was important. Though we randomized offers and nobody every declined an offer, selection could occur via the show-up rate.

Pararesult 4: The show-up rate does not vary by treatment.

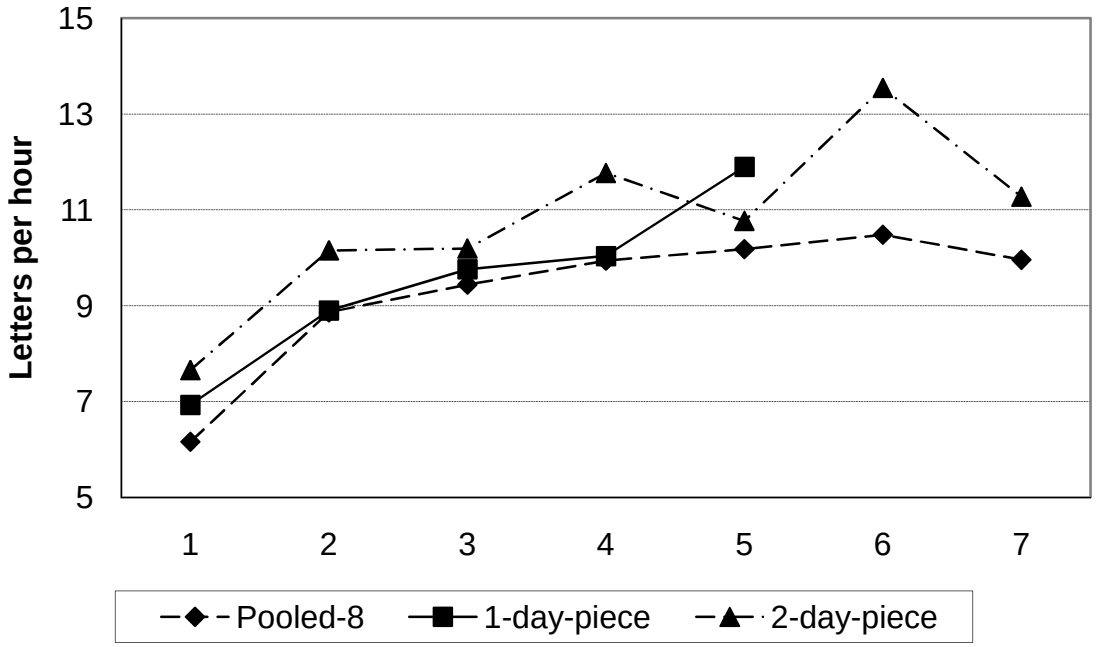
This is established by parametric (probit) estimation of the show-up rate. Thus, selection in our environment is not important.[31]

Now we turn our attention to the theoretical predictions, commencing with the pecuniary schemes. Data in Table 2 and Figure 2 are broadly in line with theoretical expectations. For example, in both piece rate treatments the number of letters produced per hour is higher than in the baseline treatment. Interestingly, the difference is a modest 6% in *1-day-piece*, whereas in *2-day-piece* the effect is much sharper, an increase of roughly 17%. Figure 3 complements these aggregate data by providing a temporal profile of the *1-day-piece* and *2-day-piece* data alongside

[31] Clearly this does not preclude selection from having an important effect in wage/productivity correlations within other worker pools (see Bewley, 1999). Recall that the natural recruiting approach for temporary workers often involves not mentioning compensation or the exact date(s) of employment on the advertisements. In terms of studying selection, this potentially differs profoundly from posting a wage and hiring the best prospects.

the fixed wage $8 data.[32] The figure provides ocular evidence that the treatment effect in *2-day-piece* is sharp across all time periods.

Figure 3: Envelopes packed by hour – piece-rate treatments



Result 1: Compared to the flat hourly baseline, there is some evidence that workers in the piece-rate produce more output at the expense of higher errors.

All of our non-parametric empirical tests reveal that the difference between *2-day-piece* and the baseline are marginally significant at conventional levels. On the other hand, the differences are not significant for *1-day-piece*. Regression evidence can be found in columns 1 and 2 of Table 3, where it is found that regressing envelopes completed per hour on the treatment dummy variables and the worker-specific observables yields similar evidence.

[32] Switching to an econometric specification where we interact the treatment dummy with the hourly dummies rather than having a single treatment dummy does not substantively alter the results. This is actually true for all the empirics that follow with one exception that we will mention.

Table 3: Regression results – piece rates

Baseline Treatment	Pooled $8 1-day-PR	Pooled $8 2-day-PR	Pooled $8 Pooled PR	Pooled $8 Pooled PR	1-day-8 Pooled PR	2-day-8 Pooled PR
Dependent variable	Letters /hour	Letters /hour	Critical errors/letter	Non-critical errors/letter	Recording errors/letter	Recording errors/letter
Constant	8.696***	11.325***	-0.010	-0.241	-0.163	-0.025
	[2.800]	[3.107]	[0.023]	[0.242]	[0.028]	[0.029]
2nd hour (dummy)	2.430***	2.632***	-0.002	0.004	0.001	0.007
	[0.458]	[0.425]	[0.006]	[0.027]	[0.028]	[0.029]
3rd hour (dummy)	3.114***	3.026***	0.005	0.001	0.004	0.032
	[0.458]	[0.425]	[0.006]	[0.027]	[0.028]	[0.029]
4th hour (dummy)	3.532***	3.895***	-0.008	-0.049*	0.006	0.028
	[0.458]	[0.425]	[0.006]	[0.027]	[0.028]	[0.029]
5th hour (dummy)	4.301***	3.711***	-0.004	-0.052*	0.027	0.032
	[0.460]	[0.425]	[0.006]	[0.027]	[0.028]	[0.029]
6th hour (dummy)	N/A	5.176***	-0.008	-0.039	N/A	0.062*
		[0.503]	[0.007]	[0.035]		[0.035]
7th hour (dummy)	N/A	3.799***	-0.008	-0.054	N/A	0.046
		[0.503]	[0.007]	[0.035]		[0.035]
Female (dummy)	-0.663	0.202	-0.002	-0.035	-0.006	-0.050
	[0.615]	[0.649]	[0.005]	[0.053]	[0.102]	[0.086]
Age in years	-0.172	-0.262	0.000	0.0271**	0.034	0.016
	[0.150]	[0.175]	[0.001]	[0.013]	[0.027]	[0.020]
Age in years squared	0.001	0.002	0.000	0.000*	0.000	0.000
	[0.002]	[0.002]	[0.000]	[0.000]	[0.000]	[0.000]
Black (dummy)	0.297	-1.900**	0.009	0.034	-0.093	-0.053
	[0.690]	[0.789]	[0.006]	[0.062]	[0.119]	[0.102]
High school maximum (dummy)	2.288***	2.651***	-0.007	-0.021	-0.089	-0.208**
	[0.701]	[0.766]	[0.006]	[0.063]	[0.122]	[0.102]
Bachelor's degree maximum (dummy)	2.632***	3.361***	0.000	-0.129*	-0.193	-0.327**
	[0.843]	[0.929]	[0.007]	[0.074]	[0.142]	[0.116]
Piece rate (dummy)	0.331	0.877	0.010**	0.015	-0.030	0.176**
	[0.527]	[0.601]	[0.004]	[0.045]	[0.092]	[0.072]
Observations	394	470	577	577	374	454
R-squared	0.256	0.333	0.052	0.102	0.040	0.164

All regressions contain random effects. Standard errors in brackets. Significance: *=10%, **=5%, ***=1%. Table 3 shows linear regressions of each dimension of productivity (column) on controls (rows) in the piece-rate treatments. See the notes in Table 2 for the definitions of each productivity and control variable. Each regression contains data from a $8/hr baseline and the piece-rate data. The variable 'piece rate' is the treatment dummy. Pooled $8 denotes a pooling of 1-day-8 and 2-day-8 data. Pooled PR denotes a pooling of 1-day-piece and 2-day-piece.

Table 3 also contains the differences in quality across treatment. Much like the quantity data, the results over quality are broadly consonant with the theoretical predictions. For example, in the pooled piece-rate schemes, the critical errors are more than triple the critical errors observed in the baseline treatments. In addition, there are perceptible differences in some of the recording errors: the pooled piece-rate data show much larger error rates compared to the *2-day-8* data. Such differences also gain statistical significance in many cases. For example, the differences in critical errors are statistically significant at conventional levels across the pooled-piece rate data and the fixed wage data using both conditional and unconditional tests. And, the recording errors are substantially larger in the pooled piece-rate data compared to the *2-day-8* data.[33]

Result 2: Compared to the flat hourly baseline, there is some evidence that the variance of workers' output and errors is higher in the piece-rate.

Using a Levene-test of change-in-variance, we find evidence that the variance of hourly worker output triples when comparing the pooled baseline to *2-day-piece* (significant at the $p < 0.01$ level), though the 1-day-piece does not gain statistical significant at conventional levels (despite increasing). In the quality dimension we see a similar pattern: the variance of critical errors more than quadruples in the pooled piece-rate compared to the baseline ($p < 0.03$ level). And, in the case of recording errors, the variance doubles and is significant ($p < 0.01$) when using *2-day-8* as the baseline. We find no difference in the variance for non-critical errors.

We now shift our attention to the non-pecuniary schemes. We utilized two different ways of inducing a positive gift. The first, *1-day-16*, was simply hiring workers at, and paying them, a higher wage.

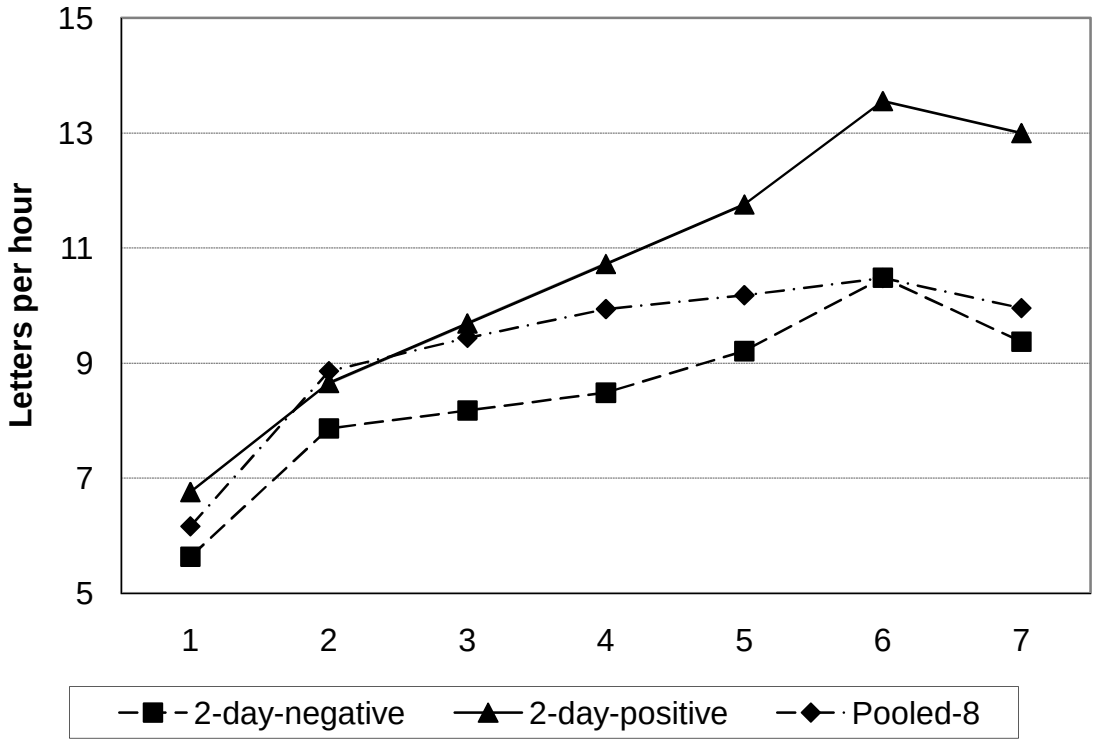
Result 3: Compared to the flat hourly market baseline, doubling hourly wages at the hiring stage has virtually no effect on output or errors.

Parametric and non-parametric tests are all insignificant at conventional significance levels. The dimension that comes closest to significance is output, which is roughly 9% higher in *1-day-16* ($p < 0.2$ using a t-test).

[33] Toilet-break data indicates insignificant differences across treatments, but on this occasion breaks seem to be longer under piece-rate, which is counterintuitive.

The second approach, *2-day-positive*, was to introduce an element of surprise by hiring workers at a wage somewhere between $8/hr-16/hr and then declaring that it would be $16/hr minutes before work commenced. Figure 4 complements the aggregate data (Table 2 and Figure 2) by providing a temporal profile of the difference between *2-day-positive* and the baseline, where the data have been averaged across individuals.

Figure 4: Envelopes packed by hour – reciprocity treatments



Result 4: Compared to the flat-hourly market baseline, there is some evidence that surprising workers with a doubling of wages leads to increased output and decreased effort, i.e., workers reciprocate positive gifts in observable effort only.

2-day-positive yielded an increase of nearly 15% in letters per hour (from 9.1 to 10.5) over the $8 fixed wage treatment. While using a Mann-Whitney non-parametric test, the difference is only statistically significant at marginal levels ($z = 1.54$; $p < 0.12$), using a two-sample t-test the difference is statistically significant at conventional levels ($t = 2.03$; $p < .05$). Yet, regressing envelopes completed per hour on the treatment dummy variables and the worker-specific observables (as well as random effects) yields much weaker statistical evidence. We present

these results in column 1 of Table 4. Here we find that upon controlling for the worker specific observables, the difference in productivity is 0.6 envelopes per hour (6.5%), and the marginal effect is significant at the $p < .28$ level using a two-tailed test.[34]

Both the number of critical and non-critical errors are less in *2-day-positive* than in the pooled baseline treatments (see Table 2), but neither is statistically significant at conventional levels. This result can be seen most clearly in columns 2 and 3 in Table 4, where the error rate is regressed on the dummy treatment variables and the worker specific observables.

One error difference that is statistically significant, however, is the number of recording errors. In this case, workers in *2-day-positive* treatment have over double the error rate compared to *2-day-8* workers (recall that we could not pool *1-day-8* data with *2-day-8* data for this type of error), and this difference is statistically significant at the $p < .08$ level. This finding is contained in column 4 of Table 4, where the recording errors are a function of dummy treatment variables and the worker specific observables.

Given that the recording task represents a worker activity that is easy to manage and with some effort higher quality can be achieved, this result provides a reason to be cautious in drawing inference concerning the quantity influence found in *2-day-positive*. The data suggest that workers are seemingly substituting their effort from outcomes that are not measured in the short run, such as proper record keeping, to short-term observable outcomes that are measured frequently, such as letters per hour. In terms of the theory laid out in section II, they are reciprocating in observables at the expense of unobservables.

[34] As is evident from the temporal profile in figure 3, the marginal significance of the quantity results for *2-day-positive* is driven by the last two hours. Running a regression with hour dummies interacting with a treatment dummy reveals that the last two are significant at the $p < 0.01$ level. As mentioned in footnote 30, this is the only regression where results from hour-by-hour treatment dummies differ substantially from a single treatment dummy.

Table 4: Regression results – positive reciprocity

Dependent variable	Letters/hour	Critical errors/envelope	Non-critical errors/envelope	Recording errors/envelope
Constant	9.018***	0.001	-0.569**	-0.407
	[3.195]	[0.012]	[0.255]	[0.446]
2nd hour (dummy)	2.405***	0.004	0.035	-0.022
	[0.395]	[0.003]	[0.026]	[0.030]
3rd hour (dummy)	3.151***	0.003	-0.014	-0.015
	[0.395]	[0.003]	[0.026]	[0.030]
4th hour (dummy)	3.848***	0.001	-0.027	0.005
	[0.395]	[0.003]	[0.026]	[0.030]
5th hour (dummy)	4.380***	-0.001	-0.013	0.035
	[0.395]	[0.003]	[0.026]	[0.030]
6th hour (dummy)	5.584***	0.000	-0.046	0.054*
	[0.457]	[0.003]	[0.030]	[0.030]
7th hour (dummy)	5.044***	0.001	-0.081***	0.066**
	[0.457]	[0.003]	[0.030]	[0.030]
Female (dummy)	0.383	-0.002	0.046***	0.024
	[0.614]	[0.002]	[0.014]	[0.093]
Age in years	-0.157	0.000	-0.001***	0.039
	[0.173]	[0.000]	[0.000]	[0.024]
Age in years squared	0.001	0.000	0.026	0.000
	[0.002]	[0.000]	[0.057]	[0.000]
Black (dummy)	-0.94	0.004	-0.087	-0.035
	[0.712]	[0.003]	[0.061]	[0.106]
High school maximum (dummy)	1.852**	-0.002	-0.087	-0.318
	[0.769]	[0.003]	[0.061]	[0.122]
Bachelor's degree maximum (dummy)	2.476***	0.000	-0.121	-0.392
	[0.941]	[0.004]	[0.075]	[0.140]
Positive reciprocity (dummy)	0.604	-0.002	-0.016	0.146*
	[0.552]	[0.002]	[0.044]	[0.074]
Observations	495	479	479	356
R-squared	0.338	0.019	0.232	0.276

All regressions contain random effects. Standard errors in brackets. Significance: *=10%, **=5%, ***=1%. Table 4 shows linear regressions of each dimension of productivity (column) on various controls (rows) in the positive reciprocity treatment, i.e., where workers are surprised by receiving higher than the market wage. See the notes in Table 2 for the definitions of each productivity and control variable. Each regression contains data from a $8/hr baseline and the positive reciprocity data. The variable 'positive reciprocity' is the treatment dummy. The $8/hr baseline is generated by pooling 1-day-8 and 2-day-8 data, with the exception of recording errors, where the baseline is only 2-day-8. See Table 1 for treatment definitions.

Comparing the quantity data across the baseline, *1-day-16*, and *2-day-positive*, we are able to provide a means to measure the pure wage effect and the surprise wage effect side by side. Interestingly, output per hour varies from 9.1 to 9.7 to 10.5 across these three treatments.

Pararesult 5: The pure wage effect is responsible for about half of the increase in output observed in the positive reciprocity treatment (baseline versus *2-day-positive*), whereas the surprise wage effect accounts for slightly more than half of the increase.

We now examine the negative reciprocity results. In *2-day-negative*, workers were hired for between \$8/hr-\$16/hr and were told that it would be \$8/hr minutes before work commenced. Figure 4 complements the aggregate data (Table 2 and Figure 3) by providing a temporal profile of the data across the baseline, *2-day-positive*, and *2-day-negative*.

Result 5: Compared to the flat-hourly market baseline, there is some evidence that surprising workers with a decrease in wages leads to less output and less effort, i.e., workers reciprocate negative gifts in both observable and unobservable effort.

Workers in the baseline treatment produce roughly 10% more output per hour than workers in the negative reciprocity treatment. Both a Mann-Whitney test and a two sample t-test indicate that this difference is statistically significant at marginal levels ($p < 0.15$ in both cases). Regressing envelopes completed per hour on the treatment dummy variables and the worker-specific observables yields similar evidence, as summarized in column 1 of Table 5. In this case, we find that upon controlling for observables the negative reciprocity condition is significant at the $p < .08$ level. In this model, workers produce 0.9 letters less per hour, a decrease of nearly 10%.

In terms of the quality dimension, negative reciprocity workers make slightly more critical and less non-critical errors than baseline workers, and a greater number of recording errors than the workers in *2-day-8*. Yet, the only dimension that gains marginal statistical significance is the recording error data, where we find that workers in the negative reciprocity treatment commit significantly more errors than baseline workers (at the $p < .09$ level in column 4 of table 5). Given the level of noise associated with the other error data, none of our statistical tests are able to achieve significance at conventional levels.

Table 5: Regression results – negative reciprocity

Dependent variable	Letters/hour	Critical errors/envelope	Non-critical errors/envelope	Recording errors/envelope
Constant	8.532***	-0.003	-0.174	-0.161
	[2.276]	[0.012]	[0.203]	[0.379]
2nd hour (dummy)	2.511***	0.000	0.018	-0.008
	[0.356]	[0.004]	[0.028]	[0.030]
3rd hour (dummy)	2.992***	-0.001	0.002	-0.028
	[0.356]	[0.004]	[0.028]	[0.030]
4th hour (dummy)	3.423***	0.001	-0.005	-0.021
	[0.356]	[0.004]	[0.028]	[0.030]
5th hour (dummy)	3.840***	0.001	0.008	-0.008
	[0.356]	[0.004]	[0.028]	[0.030]
6th hour (dummy)	4.764***	-0.005	-0.029	-0.023
	[0.412]	[0.005]	[0.032]	[0.030]
7th hour (dummy)	3.924***	-0.004	-0.032	-0.019
	[0.412]	[0.005]	[0.032]	[0.030]
Female (dummy)	-0.162	-0.005	-0.048	0.117
	[0.516]	[0.003]	[0.047]	[0.092]
Age in years	-0.111	0.001	0.020**	0.011
	[0.107]	[0.001]	[0.010]	[0.017]
Age in years squared	0.001	0.000	-0.000*	-0.000
	[0.001]	[0.000]	[0.000]	[0.000]
Black (dummy)	-0.554	0.005	0.012	0.155
	[0.689]	[0.004]	[0.062]	[0.138]
High school maximum (dummy)	1.571**	-0.007**	0.016	-0.204*
	[0.639]	[0.004]	[0.057]	[0.122]
Bachelor's degree maximum (dummy)	2.524***	-0.005	-0.054	-0.376***
	[0.779]	[0.004]	[0.070]	[0.146]
Negative reciprocity (dummy)	-0.901*	0.001	-0.059	0.144*
	[0.500]	[0.003]	[0.045]	[0.084]
Observations	496	481	481	358
R-squared	0.28	0.029	0.114	0.163

All regressions contain random effects. Standard errors in brackets. Significance: *=10%, **=5%, ***=1%. Table 4 shows linear regressions of each dimension of productivity (column) on various controls (rows) in the negative reciprocity treatment, i.e., where workers are surprised by receiving a wage lower than they expected. See the notes in Table 2 for the definitions of each productivity and control variable. Each regression contains data from a $8/hr baseline and the negative reciprocity data. The variable 'negative reciprocity' is the treatment dummy. The $8/hr baseline is generated by pooling 1-day-8 and 2-day-8 data, with the exception of recording errors, where the baseline is only 2-day-8. See Table 1 for treatment definitions.

Empirical results across the positive and negative reciprocity treatments provide an interesting confirmation of received laboratory experiments. One stylized fact in this literature is that negative reciprocity is stronger than positive reciprocity (see, e.g., Offerman (2002)). Perhaps a signal that our experimental treatments were able to invoke symmetrical negative and positive frames, we find similar insights: even in our environment where reputational concerns might seemingly swamp reciprocity effects, our results highlight the strength of negative reciprocity invoked by *2-day-negative*.

B. The effect of shocks

To examine the data from the treatments wherein we shocked the market, we use several empirical methods. Of course, with these extensions the basic problem is that we do not have a true counterfactual: the shock to the wage was done to everyone in each treatment and the baseline was itself shocked. We therefore cannot use a difference-in-difference approach, and are left with two strategies: first, we can compare pre-shock performance to post-shock performance directly. This confounds the treatment effect with a time effect. Second, under the assumption that the time effect is additively separable and constant across treatments, we can conduct a difference-in-difference using *2-day-negative* as the baseline (since this treatment continued for the full nine hours with no shock to the contract in the second day). Empirical results using all these methods point in the same direction, though the various tests vary in their levels of significance.

In sum, using both non-parametric and parametric tests, as well as regressions with individual and demographic controls, we find that the results are almost perfectly consistent with the broader treatment effects discussed above (results 1, 4 and 5). When we positively shocked income, output increases and so do two out of the three errors, with a particularly strong effect on recording errors. When income is negatively shocked, output goes down, and two out of three errors increase (the same two that have a positive treatment effect in *2-day-negative*), with recording errors again showing a large jump. When the piece-rate incentive is intensified, again output increases, along with two out of three errors, and again recording errors are the largest mover.

Finally, in an attempt to assess the cost of errors, we analyzed the results of the mailing drive. Of the approximately 5,500 envelopes mailed, only 10 recipients responded with a donation, implying a donation rate a shade under 0.2%, which is in the range of success rates for new charities using "cold" mailers. The average donation conditional on donating was $42, with a standard deviation of $34. Therefore, the average letter generates about 8¢ in donations. Clearly, using such a small sample, it is virtually impossible to make valuable statistical inference, yet it nevertheless might be informative to provide anecdotal evidence from the data.

First, given that letters with critical errors never yielded a donation, it follows immediately that a lower bound estimate of the cost of a critical error is 8¢. We consider this a lower bound estimate because once a donor gives, she is more likely to give in the future: anecdotal evidence suggests that the retention rate is roughly 50%-80% (i.e., 50%-80% of those who initially donate will contribute during the next round of solicitations). And, of those 50%-80% who are retained, they give approximately $100 per year (see Landry et al. (2006)). Second, to calculate the cost of a non-critical error, we must examine the errors committed in the envelopes that actually generated donations. We find no significant differences across letters with and without non-critical errors – one letter with a non-critical error yielded a donation. Yet, it is worth noting that in the one successful letter with a non-critical error, the error was in the color of reply card included, which should have no causal effect on the donation rate. Finally, recording errors cannot be priced since they do not affect the donation decision – they reflect an administrative task internal to the operation.

V. Conclusion

In one-shot environments, both pecuniary and non-pecuniary incentive schemes can potentially solve multitasking problems. An inherent problem with pecuniary incentive schemes, however, is that workers might substantially reduce their effort on tasks that produce unobservable outputs as they seek the salient reward to observable effort. The problems with non-pecuniary schemes are two-fold: first, if social preferences are weak, it is unlikely to be cost effective. Second, even if they are strong, if workers choose to focus their reciprocity in observable effort, unobservable effort might suffer.

This study provides empirical insights into the multitasking problem using a unique sample of workers. By simultaneously being the employer and experimenter, we were able to acquire distinctive quality data by expending an unprofitable amount on monitoring quality. Employees believe that quality is poorly monitored when it is actually precisely monitored, and so the classic multi-tasking theory applies.

Our results are broadly consonant with the theory. Workers respond to the introduction of high powered incentive schemes by increasing output at the expense of quality, and the variance of their productivity increases. Workers also respond to non-pecuniary incentives. Thus, in answer to the question: *for love or money*?, we would say unequivocally for both. Interestingly, non-pecuniary schemes generate asymmetric responses: when faced with positive gifts, workers focus their reciprocity on output at the expense of quality, yet when faced with negative gifts, all dimensions of effort deteriorate.

One overarching lesson learned from this exercise is that, as a whole, behavioral differences across the incentive schemes are much less than what one would have expected from parallel laboratory experimental results. We believe that this is chiefly a result of the intermediate reputational concerns in contrast to the one-shot frame of laboratory studies. Yet, this area is certainly ripe for future research.

At first sight, the ‘swamping’ effect of the reputational concerns may lead one to regard the environment as unsuitable for investigating pecuniary and non-pecuniary schemes. However given the prevalence of temporary workers and the near absence of genuinely one-shot employer-worker relationships, we believe that this study sheds important light on an integral component of the workforce. We trust that future research will pick up where we left off and continue to build a bridge between the laboratory and naturally-occurring environment to understand more fully how each of the factors that vary across work environments influences behavior.

References

Agell, J. and P. Lundborg (1995). "Theories of pay and unemployment: survey evidence from Swedish manufacturing firms," Scandinavian Journal of Economics. 97, p295-307.

Akerlof, G. (1982). "Labor Contracts as Partial Gift Exchange," Quarterly Journal of Economics. 97, p543-569.

American Staffing Association (2006). American Staffing 2006: Annual Economic Analysis. http://www.americanstaffing.net/statistics/economic.cfm

Baker, G. (1992). "Incentive contracts and performance measurement," Journal of Political Economy. 100, p598-614.

Baker, G., R. Gibbons and K. Murphy (1994). "Subjective performance measures in optimal incentive contracts," Quarterly Journal of Economics. 109, p1125-56.

Bewley, T. (1999). Why Wages Don't Fall During a Recession. Harvard University Press, Cambridge, Massachusetts.

Charness, G. (2005). "Attribution and reciprocity in an experimental labor market," Journal of Labor Economics. 22(3), p665-88.

Charness, G., G. Frechette and J. Kagel (2004). "How robust is laboratory gift exchange?," Experimental Economics. forthcoming.

Chen, P. (2005). "Reciprocity at the workplace: do fair wages lead to higher effort, productivity, and profitability?," Mimeo, Australian National University.

Dufwenberg, M. and G. Kirchsteiger (2004). "A theory of sequential reciprocity," Games and Economic Behavior. 47, p268-98.

Engelmann, D. and A. Ortmann (2002). "The robustness of laboratory gift exchange: a reconsideration," working paper, Charles University.

Falk, A., and U. Fischbacher (2006). "A theory of reciprocity," Games and Economic Behavior. 54, p293-315.

Fehr, E., G. Kirchsteiger and A.Riedl (1993) "Does fairness prevent market clearing? An experimental investigation," Quarterly Journal of Economics. 108(2)m p437-60.

Fehr, E., S. Gächter and G. Kirchsteiger (1997). "Reciprocity as a contract enforcement device," Econometrica. 65(4), p833-60.

Gneezy, U. and J. List (2006). "Putting Behavioral Economics to Work: Field Evidence of Gift Exchange," Econometrica, September, 74(5), p1365-1384.

Hennig-Schmidt, H., B. Rockenbach and A. Sadrieh (2005). "In search of workers' real effort reciprocity – a field experiment and a laboratory experiment," Mimeo, University of Bonn.

Holmstrom, B. and P. Milgrom (1991). "Multitask principal-agent analyses: incentive contracts, asset ownership and job design," Journal of Law, Economics and Organization. 7, p24-52.

Johnson, R., D. Reiley and J. Munoz (2006). "The war for the fare: how driver compensation affects bus system performance," Working paper, University of Arizona.

Katz, L. (1986). “Efficiency Wage Theories: A Partial Evaluation,” in NBER Macroeconomics Annual, ed. S. Fischer, Cambridge, MA: MIT Press.

Krueger, A. and A. Mas (2004). “Strikes, scabs and tread separations,” Journal of Political Economy. 112(2), p253-89.

Kube, S., M. Marechal and C. Puppe (2006). “Putting reciprocity to work – positive versus negative responses in the field,” Mimeo, University of Karlsruhe.

Landry, C., A. Lange, J.A. List, M.K. Price, and Nicholas Rupp. 2006. "Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment,” Quarterly Journal of Economics, 121 (2): 747-782.

Lazear, E. (2000). “Performance pay and productivity,” American Economic Review. 90(5), p1346-61.

Lee, D. and N. Rupp (2006). “Retracting a gift: how does employee effort respond to wage reductions?,” Mimeo, East Carolina University.

Mas, A. (2005). “Pay, reference points and police performance,” Mimeo, University of California, Berkeley.

Offerman, T. (2002). “Hurting hurts more than helping helps,” European Economic Review. 46, p1423-37.

Paarsch, H. and B. Shearer (2000). “Piece rates, fixed wages and incentives effects: statistical evidence from payroll records,” International Economic Review. 41(1), p59-92.

Pritchard, R., M. Dunnette and D. Jorgensen. (1972). “Effects of perceptions of equity and inequity on worker performance and satisfaction,” Journal of Applied Psychology. 56(1), p75-94.

Shearer, B. (2004). “Piece rates, fixed wages and incentives: evidence from a field experiment,” Review of Economic Studies. 71, p513-34.