The effect of intrinsic motivation and reference point updating on effort provision: a laboratory experiment

Giovanni Giusti

September 2014

Abstract

Despite the fact that much research has been devoted to understanding the effect of performance-contingent monetary incentives on individuals' effort provision, little attention has been dedicated toward understanding whether a sequence of changes in the amount of piece rate an individual receives affects performance.

In this paper we examine how effort provision of participants varies in a three-period laboratory setting in which different performance based compensation schemes are implemented for two differentiated tasks.

The first task, consisting of downward counting from a big number while each time subtracting a fixed quantity, is considered relatively more challenging, while the second task consisting of counting the number of A's in a paragraph is considered relatively less challenging.

Furthermore, tasks differed on subjects' individual evaluation on an interest/enjoyment scale.

Our findings are as follows: (1) Performance contingent incentives affect participants' effort provision. (2) The effect is task dependent and it is much stronger for the less challenging task. (3) A unique increase in the amount of piece rate leads to an increase in performance only in the less challenging task. (4) A decrease in piece rate incentive negatively affects subjects' performance on both tasks, but only provided that the decrease follows a previous increase.

1. Introduction

Consider the situation in which the amount of piece rate payment a worker receives for his/her job changes through time in a non-performance contingent way.

This can be quite common, especially for some types of low level seasonal jobs including fruit pickers or tree-planters.

In these types of jobs, a worker is often paid for example based either on the number of boxes he/she manages to fill each day or the number of trees he/she plants. Thus, he/she receives a salary that depends on the level of effort he/she exerts.

Often, the amount of piece rate that the employer pays changes substantially from season to season, on a monthly base, or from one crop to another.

There can be several reasons for these changes. For example, the market price at which a specific good is sold is lower this season compared to the previous one, or because planting conditions vary substantially among soils (Paarsch and Shearer, 2009).

Considering the high workforce mobility in these types of jobs, it could also be a case of the worker being aware that in the previous seasons a specific plantation was paying a different piece rate amount.

Thus, even if that worker was not employed there previously, he/she might arrive with some expectations formed which differ from the reality.

We believe that an important issue is understanding if these changes in piece rate from a previous level can affect worker's motivation and, consequently, the level of effort he/she is willing to exert.

The answer to such questions should be of particular interest to an employer who is looking for the best available contract that maximizes firm's profit.

Despite the ease of finding data regarding employees' piece rate amounts received and their performance, establishing a causal relationship is very hard. This is because there are many unobservable factors that may affect this relationship.

The best way to control for these confounding factors and to eliminate endogeneity problems is to study the question in an environment in which changes in piece rate are completely exogenous.

For this reason, we decided to study how changes in the amount of piece rate incentive affect task performance by using a laboratory experiment.

In a three period laboratory experiment we implemented four different incentive schemes across three distinct periods.

(1) Fixed payment: participants received a constant and fixed amount of money in each period that did not depend on the number of correct or incorrect answers.

(2) Constant piece rate[1]*:* participants received a piece rate per correct answer and received a penalty for incorrect answers. The amount remained unchanged over the three periods.

(3) Peak piece rate: The piece rate received (and penalty) is doubled in the second period compared to the first and third periods.

(4) Valley piece rate: The piece rate received (and penalty) is halved in the second period compared to the first and third periods.

Each treatment is implemented in two different tasks. The first task was designed to be relatively more challenging while the second task was designed to be less challenging. This task difference was confirmed by subjects' average evaluations on a post-treatment interest/enjoyment scale.

We find that, consistent with standard economic theory, performance contingent monetary incentives affect subjects' effort provision. At the same time, as suggested by the literature on intrinsic motivation, we found that the effect of monetary incentives varies between tasks and is much stronger in the less challenging task.

We also obtained that a unique, positive change in piece rate leads to an increase in subjects' performance, but only in the less challenging task.

We also obtained that a unique, positive change in piece rate leads to an increase in subjects' performance, but only in the less challenging task.

Finally, in line with literature based on reference-dependent outcomes, we discovered that a decrease in piece rate that follows an increase negatively affects performance on both tasks, while an increase that follows a decrease does not significantly affect performance.

The main message that we have derived from our results is that employers of especially boring and repetitive types of jobs should not increase the amount of piece rate if they cannot maintain the increase over time.

[1] p.r. is often used in the future as abbreviation for the words "piece rate".

2. Literature review

In this section we will refer to four different but related literatures from which we derived our set of hypotheses.

2.1 Effect of monetary incentive on performance

In economic theory, incentives serve as an instrument to align principal-agent objectives (Holmstrom, 1979; Kale et al., 2009; Gibbons, 1992).

For an employer, it is therefore a common practice to use financial incentives to improve employees' performance and thus increase the firm's profit.

Several papers support this view.

Lazear (2000), for instance, found that the average level of output per worker increases between 20% and 36% when switching from hourly wage rate to performance based payment.

Banker et al. (1996) analyzed a panel of 15 retail outlets and found that sales increased when performance-based compensation was introduced.

In a field experiment, Hossain and List (2009) found that incentives lead to higher productivity, especially for teams.

In laboratory experiments, Libby et al. (1992) found that subjects worked harder and were more accurate on a recall and recognition task if rewarded for performance. Fehr, Goette and Lienhard (2008) showed that effort provision of participants incremented by 15% when piece rate substituted fixed wages.

Despite much evidence, several other papers found that under some conditions, or in some specific tasks, monetary incentives do not significantly affect workers' performance (see Camerer and Hogarth, 1999 and Bonner and Sprinkle, 2002 for a comprehensive review of the effect of incentive in performance).

2.2 Incentive and intrinsic motivation

One situation in which the presence of monetary incentives might affect performance differently from what economic theory predicts is, for instance, when workers are intrinsically motivated by doing the task (Frey, 1997).

According to Deci (1971) "One is said to be intrinsically motivated to perform an activity when he receives no apparent reward except the activity itself" (p.105).

In a famous experiment, *Deci*, Koestner, & *Ryan*, (1999) showed that adding monetary incentives (i.e. extrinsic reward) for an intrinsically motivated task can "crowd out" motivation and consequently undermine individuals' effort.

But most of the time extrinsic incentives co-exist with intrinsic motivation, given that a person could enjoy doing a task and at the same time be paid for doing it.

Therefore, recent research has tried to assess how the presence of intrinsic motivation interacts with the presence of monetary incentives and which is the most efficient combination between them (Cerasoli et al, 2014).

Results suggest that monetary (i.e. extrinsic) incentives are generally more powerful for very repetitive and boring tasks since they are characterized by low levels of intrinsic motivation (Cerasoli et al, 2014).

2.3 Incentive magnitude and incentive variations

There are two other questions that have received attention in the incentives literature. The first investigates how *absolute magnitude* of monetary reward affects performance, and the second investigates what people's responses are to *relative changes in the magnitude of incentives*.

Regarding *absolute magnitude,* laboratory evidence showed mixed results.

For instance, Gneezy and Rustichini (2000) found that the effect of monetary compensation on performance does not monotonically increase as one might think.

In fact, they obtained that for an IQ test task, subjects who were paid a small amount of money for a correct answer performed poorer than subjects that were not paid at all. In order for monetary incentives to matter they observed that the magnitude had to be sufficiently high.

On the other hand, Pokorny (2008) and Takahashi et al. (2014) found the opposite phenomenon. For example, Pokorny (2008) found that in a laboratory experiment subjects exerted more effort when they received very low incentives compared to when they were not paid or when they were paid a lot.

With respect to the effect of *relative changes in the magnitude of incentives*, results are less mixed and seem to converge in one direction.

For instance, in a field experiment Paarsch and Shearer (2009) analyzed how positive changes in piece rate amount affect the performance of tree-planters. They observed that an increase in piece rate leads to a significant increase in performance.

Dikinson (1999) conducted a laboratory experiment and was able to disentangle income from substitution effects when varying wages. For half of the subjects he substantially increased the piece rate amount from the first to the second period, while doing the opposite for the other half. He found that when subjects could only vary their amount of effort and not the duration of work, a compensated wage increase (decrease) made subjects work harder (less hard).

2.4 Prospect theory and reference point updating

The results obtained by Dikinson (1999) and Paarsch and Shearer (1999) are very much in line with what standard economic theory predicts.

At the same time, starting with the work of Camerer et al. (1997), several empirical studies found evidence of negative temporal substitution in response to temporary wage shocks. For instance, Camerer et al. (1997) found that taxi drivers reduced the number of working hours following an increase in their per-hour salary. These results clearly go against standard economic theory and could be explained with behavioral models based on prospect theory (Kahneman and Tversky, 1979).

According to prospect theory, individuals' preferences depend not only on the achieved income level but also on some relative reference level (or reference point).

In the case of taxi drivers for instance, as pointed out by Camerer et al. (1997), the reference level was provided by the targeted income a taxi driver made on an average day. As a result, he/she would diminish his effort by reducing the number of hours worked once that level was attained.

Both prospect theory and prospect theory based reference dependence models (e.g. Koszegi and Rabin, 2006) make clear predictions regarding individuals' outcome evaluation from a reference point.

They furthermore suggest plausible candidates as reference points such as individuals' status quo or individuals' expectations.

Despite that, they do not say much regarding how reference points are updated, shift and evolve over time.

Recent research attempts to fill this gap (Baucells et al. 2011; Arkes et al. 2008, 2010; Chen and Rao 2002).

For instance, Chen and Rao (2002) proposed a theoretical model and a set of laboratory experiments to understand how people's reference points are sequentially updated.

In their work they show that the order in which reference points changes can critically affect their final evaluation.

More specifically, they studied the question of how individuals feel after being exposed to a sequence of two events of equal magnitude but opposite direction (i.e. gain followed by a loss vs. a loss followed by a gain).

They show, both theoretically and empirically, that the order in which two equivalent events occurs matters. In fact, they found that a gain followed by a loss will be evaluated less favorably (i.e. generate a higher loss in individual utility) compared to a loss followed by a gain[2].

According to them, the reason for the difference in the evaluation is due to the way people update their reference points.

In particular, they point out that a first outcome moves the reference point in a particular direction. Thus, a sequence of events, in which an initial event is surprisingly reversed, will be more favorable if the first event is a loss than if it is a gain.

3. Description of the tasks

As we have seen in the literature reviewed in section 2.2, the effect of monetary incentives critically interacts with the degree of intrinsic motivation that a task generates. For this reason it becomes particularly important for our study to control whether intrinsic motivation also affects subjects' responses to different dynamics of incentives.

Therefore, we implemented the same set of four treatments on two differentiated tasks.

The first task is assumed to be more challenging. Subjects should derive relatively more pleasure from the task itself and should, therefore, be relatively more intrinsically motivated by doing it.

The second task is assumed to be relatively less challenging and more boring. Therefore, subjects should derive relatively less intrinsic motivation from doing it.

3.1 The more challenging task

From the experimental literature in social sciences, we observe that when the experimenter intended to assign subjects a challenging and effortful task he/she often proposed exercises having subjects perform non-trivial mathematical operations.

[2] This result finds additional support in Baucells et al. 2011.

This is, for instance, the case of Brüggen and Strobel (2007) who had subjects multiply sequences of two digit numbers; or, for example, in Niederle and Vesterlund (2007) where subjects had to add sets of 4 two-digit numbers.

In our experiment, we construct a mathematical task that consists of downward counting from a very large number (e.g. 1,500,000) while continually subtracting a fixed quantity from the previous number.[3]

In particular, the numbers to be subtracted were 13, 17 and 27 respectively for the first, second and third periods. [4]

3.2 The less challenging task

In order for a task to be effortful and yet not constitute a challenge for subjects, it needs to be sufficiently boring, meaningless and repetitive.

From the experimental literature in social science we find that typical tasks with these characteristics involve, for instance, counting quantities of given numbers or letters within some sets. For example Abeler et al. (2009) made subjects count the number of zeros contained in a matrix consisting of 1s and zeros, or for instance Rey-Biel et al. (2013) made subjects count a specific letter within a fixed sequence of sentences.

Our task is similar to the one proposed by Rey-Biel et al (2013). Subjects had to count the number of A's contained in paragraphs where, in addition, the font was particularly small[5]

In our experiment, the exact same task was carried out in the three periods, each period lasting 4 minutes[6].

3.3 Post –treatment task evaluation

In order to insure that tasks differed on the level of intrinsic motivation they generated, we made subjects rate them by answering a post treatment questionnaire.

Specifically, we used the interest/enjoyment subscale taken from the Intrinsic Motivation Inventory scale (Ryan, 1982). This subscale is considered to be the self-

---

[3] We could also have used the same task as Brüggen and Strobel (2007) or as Niederle and Vesterlund (2007). At the same time our task is conceived to be relatively more entertaining and for this reason might perhaps induce a relatively higher level of intrinsic motivation (see footnote 4).

[4] In a previous pilot session (with 15 paid subjects) we individually asked (with open type of questions) to the subjects their perception about the task. Their answers suggested that they perceived the task as challenging and entertaining.

[5] A screenshot of both tasks is reported in the appendix with its description.

[6] Experimental sessions for the boring task were always run at the end of another experiment of an unrelated subject.

reported measure of intrinsic motivation and has been extensively used, especially in the cognitive psychology literature (e.g. Ryan, 1982; Ryan, Mims & Koestner, 1983).

The interest/enjoyment subscale is composed of seven questions below. Each question was rated from 1 to 5. 1 = Not at all true, 2 = Not true, 3 = Somewhat true, 4 = True and 5 = Very true.

The questions were as follows[7]:

1. I enjoyed doing this activity very much
2. This activity was fun to do
3. I thought this was a boring activity
4. This activity did not hold my attention at all
5. I would describe this activity as very interesting
6. I thought this activity was quite enjoyable
7. While I was doing this activity, I was thinking about how much I enjoyed it

4. Experimental design

The experiment took place at the Leex (Laboratory of Experimental Economics) at the University Pompeu Fabra between spring 2010 and spring 2011. Participants were recruited via the Leex list of participants which is mostly composed of undergraduate students from all discipline. Participants' average age was of 21 and the number of males was approximately the same as the number of females.

The experiment consisted of eight treatments. The same four treatments were implemented in two different tasks in a 2x4 between subjects design. Tasks were programmed with Ztree (Fischbacher, 2007).

Each session was composed of 20 participants. Subjects recruited in each session were randomly assigned to one of the eight treatments (four treatments for each task).

Before starting each session the experimenter read aloud the general part of the instructions, this was common to all groups for the same task. After general instructions, subjects could familiarize themselves with the task through a trial period of thirty seconds.

Before beginning each period, in each subject's screen an individual text with instructions appeared. In this screen the critical information that changes between

[7] Note that at the moment of calculating the average, the scale for questions 3 and 4 were reversed.

groups was the amount of piece rate participants received for each correct answer (and penalty for incorrect ones) for the period that was about to start.

Importantly, the change in piece rate was made particularly salient to subjects. For instance, in addition to reporting the amount of incentive in points, we used sentences such as: “Be careful! For this period the amount you earn per correct answer is doubled (or halved) compared to the previous period”.

Furthermore participants also had to write the amount of piece rate per correct answer before starting each period in a piece of paper to return at the end of the experiment[8].

In each treatment, participants did not know that the amount of piece rate would change in following periods. They simply received the new information at the beginning of each period.

Treatment	**Period 1**	**Period 2**	**Period 3**
(1) Fixed payment	Fixed payment	Fixed payment	Fixed payment
(2) Constant p.r.	P* (correct answers) – (0.25*P* incorrect answers)	P* (correct answers) – (0.25*P* incorrect answers)	P* (number of correct answers) – (0.25*P* incorrect answers)
(3) Peak p.r.	P* (correct answers) – (0.25*P* incorrect answers)	**2P*** (number of correct answers) – (**0.5***P* incorrect answers)	P* (number of correct answers) – (0.25*P* incorrect answers))
(4)Valley p.r.	P* (correct answers) – (0.25*P* incorrect answers)	½ **P*** (number of correct answers) – (**0.125***P* incorrect answers)	P* (number of correct answers) – (0.25*P* incorrect answers)

Table 1: Experimental structure

Table 1 reports the characteristics of each experimental treatment for each of the two tasks.

It is important to notice that in each period of each treatment the penalty for an incorrect answer was valued only ¼ of the value of each correct answer. Given this difference we decided to carry out separately the analysis of correct and incorrect answers.[9]

[8] The reason for these two additional measures was to raise subjects’ awareness that there was a change in the piece rate they received. We decided to adopt these measures after running the pilot session given that several subjects at the end of the pilot could not remember the piece rate they received in each period.
[9] Analysis with incorrect answers is reported in section 6.2.

In the *Fixed payment* subjects received 3 Euros in each period of the more challenging task and 2 Euros in each period of the less challenging task independently from their performance. This difference was because a period in the more challenging task lasted 8 minutes while in the less challenging task it lasted 5 minutes.
In the other treatments, payoffs depended on performance.
The piece rate for a correct answer is represented in table 1 by P which is always 100 points. What changes is the factor that multiplies P which characterized the dynamic of the incentives studied.
The exchange rate was 100 points = 4 cents for the more challenging task and 100 points = 10 cents for the less challenging task. In this way we keep the same structure of points for both tasks[10]. Average payment per subject at the end of the experiment was 8 Euros[11] for the more challenging task and 5 Euros for the less challenging task.

5. Hypotheses

Based on the literature review we developed in section 2, we present the following set of hypotheses.
In accord with standard economic theory and based on the literature we reported in section 2.1 we form the first hypothesis.

Hypothesis I: If subjects are paid based on performance, they will exert more effort compared to the case in which their payment is fixed and independent of performance.
We should therefore find that for period 1 subjects in the fixed payment treatment obtain less correct answers than subjects in other treatments[12]:

$$Fixed\ payment_1 < Average(Constant\ p.r._1 + Peak\ p.r._1 + Valley\ p.r._1)$$

Where $p.r._1$ means piece rate in period 1.

Based on the literature on the effect of relative changes in piece rate incentive magnitude described in section 2.3 we form the second hypothesis.

[10] Piece rate was set in a way that subjects' average payment with respect to the time spent in the laboratory would not differ substantially between the two tasks.
[11] In each treatment subjects received as well an additional payment of 2€ for showing up.
[12] Given that the payment per correct answer in the first period is the same for each of the piece rate treatment (i.e. 100 points), we can pool together subjects' number of correct answers.

Hypothesis II: Given the impossibility of intertemporal substitution in our experiment, a change in the amount of piece rate paid will lead to a change in the level of effort provided in the same direction of the change.

Therefore we should find that the difference between number of correct answers in period 2 and number of correct answers in period 1 are as follow[13]:

$$Peak\ p.r._{2-1} > Constant\ p.r._{2-1}$$

$$Valley\ p.r._{2-1} < Constant\ p.r._{2-1}$$

Based on the literature on asymmetric reference point updating described in section 2.4 and assuming that changes in piece rate incentive do shift reference points we form the following hypothesis for the number of correct answer in period 3:

Hypothesis III: A sequence of two consecutive changes in reference points will have a negative and stronger effect on performance when the first change is positive and the second change is negative, than the other way around.

This implies that we should find an asymmetric response to a second change in the magnitude of incentives.

Specifically we should obtain that the negative change in period 3 for the peak piece rate treatment will have a stronger effect on subjects' performance than the change in period 3 for the valley piece rate treatment.

$$Effect(Peak\ p.r._{3-2} - Constant\ p.r._{3-2})$$

$$>$$

$$Effect(Valley\ p.r._{3-2} - Constant\ p.r._{3-2})$$

[13] We use the differences in correct answers from one period to the other since we are interested in measuring performance's changes. The sub-index denotes for which periods we are calculating the difference of correct answers. For example, if the sub-index is 2-1 it means correct answers in period 2 minus correct answers in period 1.

Based on the literature on the interaction of the type of task with performance based monetary incentives reported in section 2.2 we form the fourth hypothesis.

Hypothesis IV: The effect of performance based monetary incentive on subjects' effort level will be stronger for the less challenging task.

This because in the less challenging task the level of intrinsic motivation should be lower compared to the more challenging task and consequently the overall response to extrinsic monetary incentive should be stronger.

$$Piece\ rate\ incentive\ \ Effect\ (Less\ challenging\ task)$$

$$>$$

$$Piece\ rate\ incentive\ Effect\ (More\ challenging\ task)$$

6. Experimental results

In table 1 and table 2 we report the mean and standard deviation for correct and incorrect answers for each of the three periods of the four treatments for the more challenging task and for the less challenging task, respectively. In figure 2 and 3 we report graphically the evolution of the number of correct answers for each task.

More Challenging Task

Treatment		Mean Correct	S.d. Correct	Mean Incorrect	S.d. Incorrect	N
Period 1	Fix payment	42	17	3	4	19
	Constant piece rate	47	18	2	2	20
	Peak piece rate	38	26	4	3	19
	Valley piece rate	48	14	2	3	20
Period 2	Fix payment	37	15	5	6	19
	Constant piece rate	44	16	3	3	20
	Peak piece rate	38	19	7	5	19
	Valley piece rate	43	13	3	5	20
Period 3	Fix payment	30	13	4	6	19
	Constant piece rate	38	17	3	4	20
	Peak piece rate	29	17	4	2	19
	Valley piece rate	35	15	3	3	20

Table 2: Mean and standard deviation of correct and incorrect answers - more challenging task

Less Challenging Task

Treatment		Mean Correct	S.d. Correct	Mean Incorrect	S.d. Incorrect	N
Period 1	Fix payment	3	3	3	1	20
	Constant piece rate	11	2	2	1	30
	Peak piece rate	9	3	2	2	30
	Valley piece rate	9	2	2	1	30
Period 2	Fix payment	3	2	4	2	20
	Constant piece rate	8	2	3	1	30
	Peak piece rate	12	3	3	1	30
	Valley piece rate	6	3	2	1	30
Period 3	Fix payment	5	2	4	2	20
	Constant piece rate	8	2	3	1	30
	Peak piece rate	5	3	4	1	30
	Valley piece rate	8	2	3	1	30

Table 3: Mean and standard deviation of correct and incorrect answers - less challenging task

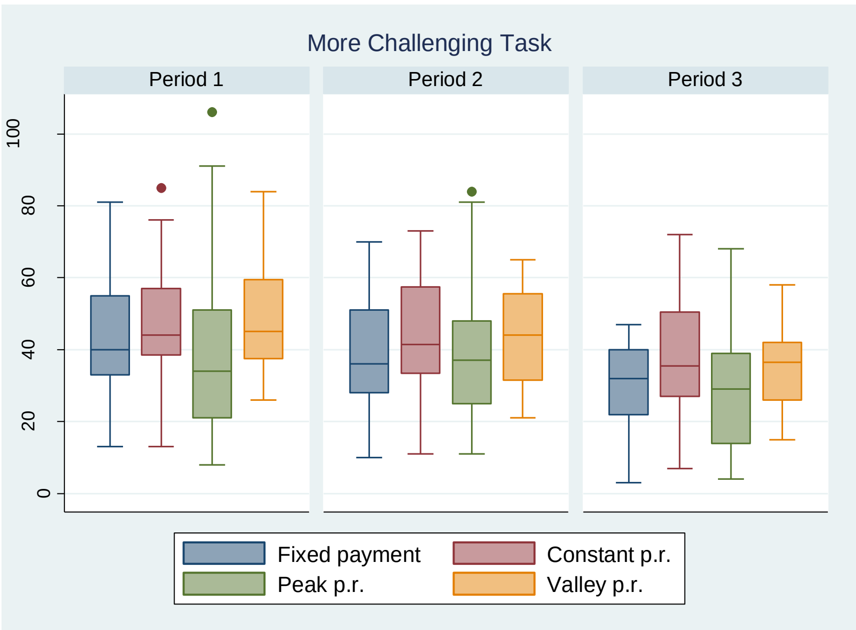


Figure 1: Number of correct answers for each period in each treatment - more challenging task

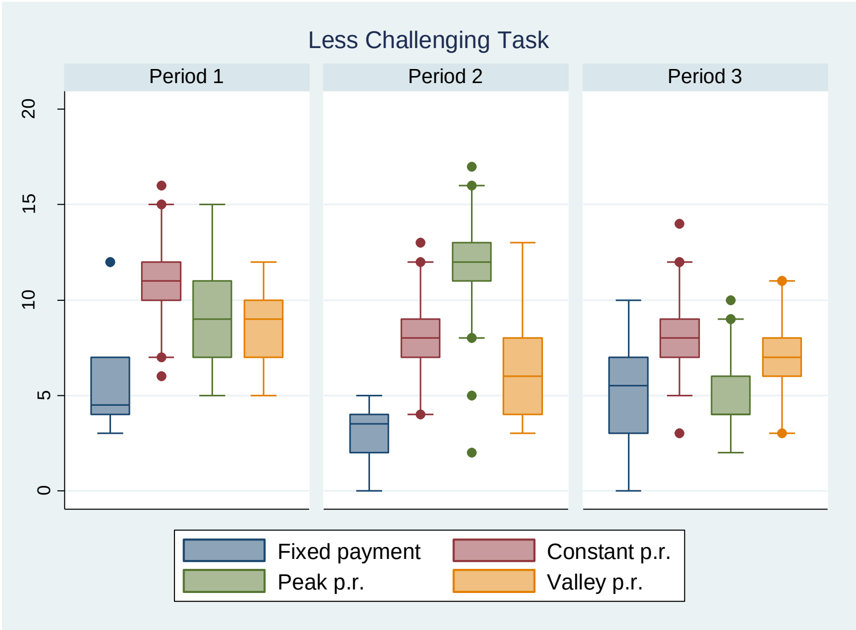


Figure 2: Number of correct answers for each period in each treatment - less challenging task

Figure 1 and figure 2 report a visual representation (i.e. a box plot) of the distribution of the number of correct answers for each of the treatments for the more challenging task and for the less challenging task respectively.

The box plot splits the distribution of correct answers into quartiles. The horizontal line contained in each box represents the median of the distribution of correct answers for each treatment for the corresponding period. The two horizontal lines outside each box represent lowest and highest observations. Points represent outliers.

From the distribution of correct answers for the more challenging task (figure 1) we observe a general tendency of the median of correct answers to decrease from period 1 to period 3 in each treatment. This could be caused both because of subjects' fatigue and because the number to be subtracted differed from period to period, perhaps increasing task difficulty in late periods.

In the same figure we also observe that within each period, medians of correct answers between treatments do not differ remarkably.

It is particularly interesting to notice that for the fixed payment treatment, median of correct answers is very similar to the median of the other three performance contingent treatments.[14]

From the distribution of correct answers for the less challenging task reported in figure 2 we observe a more marked difference between treatments.

In particular the fixed payment treatment is characterized by a much lower number of correct answers compared to each of the other treatments.

Particularly low is as well the number of correct answers in period 3 for the peak p.r. treatment and the visually substantial decrease that the peak p.r. treatment experienced from period 2 to period 3.

6.1 Hypothesis testing

The statistical inference that we present in next section will use only the number of correct answers as measure for performance[15].

The analysis of the evolution of the incorrect answers is reported separately in section 6.2.

Testing Hypothesis I

The first hypothesis suggests that for period 1 the level of effort exerted by subjects will be higher when subjects are remunerated in a performance contingent way. That is:

$$Fixed\ payment_1 < Average(Constant\ p.r._1 + Peak\ p.r._1 + Valley\ p.r._1)$$

[14] This could perhaps be explained by the fact that subjects feel particularly intrinsically motivated with this task or can also be explain by the incompleteness of contract between subjects and experimenter as speculated by Gneezy and Rustichini (2000).

[15] In order to test our hypotheses we adopt a statistical approach that is widely used in psychology and begin to be commonly used as well in other social sciences. The approach is part of what is named "*The new statistics"* (Cumming, 2013) and it suggests that instead of using the null – hypothesis significant testing approach, we should estimate our results in terms of effect sizes and confidence intervals.

The reason we favor this approach is that it allows us not only to establish whether there are differences among treated groups but instead it tells us the size of the differences and as consequence it allows us to easily compare results across different design or, as in our case across different tasks.

Furthermore, in addition to reporting effect sizes (estimated in our case by using *Cohen's d*) we also report the 95% confidence interval of each effect size. In this way we are able to capture better the extent of uncertainty in our inference.

More Challenging Task: Correct answers in period 1				
Treatment	Statistic			
	Mean	S.d.	Average p.r.	Fixed Payment
Average p.r.	44.86	17.7	------	0.13 [-0.39; 0.64]
Fixed Payment	42.36	20.24	------	------

Table 4: Testing hypothesis I - more challenging task

Less Challenging Task: Correct answers in period 1				
Treatment	Statistic			
	Mean	S.d.	Average p.r.	Fixed Payment
Average p.r.	9.72	4.26	------	1.68 [1.14; 2.21]
Fixed Payment	5.5	9.2	------	------

Table 5: Testing hypothesis I - less challenging task

Table 4 and 5 report the difference in means for the number of correct answers in period 1 in the fixed payment treatment compared with the average of the other three groups for the more challenging task and for the less challenging task respectively. Numbers reported in the right part of the table represent Cohen’s d measuring the effect size of groups’ differences and its 95% confidence interval.

Result I: The number of correct answers for the less challenging task is higher if subjects are paid for performance compared to when they received a fixed payment.

Support for result I:

From table 5 we observe a large effect size for the average number of correct answers in period 1 for the three performance contingent treatments pooled. In particular Cohen’s d is 1.68, 95% CI [1.14; 2.21] compared to the fixed payment treatment.

The same analysis for the more challenging task (table 4) shows only a very small effect size of 0.13 with a 95% CI of [-0.39; 0.64].

Testing Hypothesis II

The second hypothesis suggests that a change in the amount of piece rate paid to subjects will lead to a change in the level of effort provided in the same direction of the change.

$$Peak\ p.r._{2-1} > Constant\ p.r._{2-1}$$

$$Valley\ p.r._{2-1} < Constant\ p.r._{2-1}$$

More Challenging Task: Correct answers in period 2 - Correct answers in period 1

Treatment	**Statistic**		**Pairwise comparisons**			
	Mean	S.d.	Fixed payment	Constant p.r.	Peak p.r.	Valley p.r.
Fixed payment	-4.9	7.9	------	-0.21	-0.55	0.01
Constant p.r.	-2.9	10.26	------	------	-0.29	0.20
Peak p.r.	0.11	9.97	------	------	------	0.50
Valley p.r.	-5.00	10.29	------	------	------	------

Table 6: Testing hypothesis II - more challenging task

Less Challenging Task: Correct answers in period 2 - Correct answers in period 1

Treatment	**Statistic**		**Pairwise comparisons**			
	Mean	S.d.	Fixed payment	Constant p.r.	Peak p.r.	Valley p.r.
Fixed payment	-2.40	3.50	------	-0.23	-1.21	0.02
Constant p.r.	-3.10	2.50	------	------	-1.59	0.23
Peak p.r.	2.13	3.90	------	------	------	1.34
Valley p.r.	-2.46	2.89	------	------	------	------

Table 7: Testing hypothesis II - less challenging task

.

Tables 6 and 7 report the difference in means of correct answers given in period 2 minus the correct answers given in period 1 for the more challenging task and for the less challenging task respectively. Numbers reported in the right part of the table represent Cohen’s d and measure the effect size of groups’ differences.

Result II: A positive change in the magnitude of piece rate per correct answer leads to a positive change in subjects’ performance. Furthermore a negative change in the magnitude of piece rate per correct answer does not affect subjects’ performance.

Support for result II:

We compare mean differences between the number of correct answers in period 2 minus the correct answers in period 1for both tasks.[16]

The effect size is particularly marked once comparing the *constant p.r.* treatment (mean = -3.10) with the *peak p.r.* treatment (mean = 2.13) for the less challenging task.

The estimated differences between the means is -1.59, 95% CI [-2.17; -1.00].

The same comparison for the more challenging task also suggest a similar result although the effect size is much smaller and equal to -0.55, 95% CI [-1.20; 0.09].

In both tasks, when comparing valley p.r. and constant p.r. we find very small effect sizes. This result suggests that a decrease in piece rate per correct answer does not have a substantial effect on subjects' performance.

Testing Hypothesis III

The third hypothesis suggests that the order of the sequence in which the piece rate incentive changes matters for subjects' performance. In particular it suggests that the negative effect will be much stronger when the first change is positive and the second change is negative.

$$Effect(Peak\ p.r._{3-2} - Constant\ p.r._{3-2})$$

$$>$$

$$Effect(Valley\ p.r._{3-2} - Constant\ p.r._{3-2})$$

[16] Numbers in the right part of table 6 – 7 reports *Cohen's d*, a measure for the effect size of the difference in means between treatments. In each cell (of each tables in this paper), a negative (positive) number means that the treatment reported in the column is characterized by a highest (lowest) average compared to the treatment reported in the row of the table.

More Challenging Task: Correct answers in period 3 - Correct answers in period 2						
Treatment	**Statistic**		**Pairwise comparisons**			
	Mean	S.d.	Fixed payment	Constant p.r.	Peak p.r.	Valley p.r.
Fixed payment	-7.42	7.09	------	-0.18	0.38	0.15
Constant p.r.	-6.15	6.88	------	------	0.57	0.33
Peak p.r.	-10.05	6.79	------	------	------	-0.22
Valley p.r.	-8.50	7.15	------	------	------	------

Table 8: Testing hypothesis III - more challenging task

Less Challenging Task : Correct answers in period 3 - Correct answers in period 2						
Treatment	**Statistic**		**Pairwise comparisons**			
	Mean	S.d.	Fixed payment	Constant p.r.	Peak p.r.	Valley p.r.
Fixed payment	1.80	2.37	------	0.50	2.49	0.24
Constant p.r.	0.33	3.18	------	------	1.95	-0.26
Peak p.r.	-6.66	3.93	------	------	------	-2.24
Valley p.r.	1.13	2.95	------	------	------	------

Table 9: Testing hypothesis III - less challenging task

Table 8 and 9 report the difference in means of correct answers given in period 3 minus correct answers given in period 2 for the more challenging task and for the less challenging task respectively. Numbers reported in the right part of the table represent Cohen's d and measure the effect size of groups' differences.

Result 3: The order of the sequence in which a decrease in piece rate follows a previous increase made subjects to reduce substantially the number of correct answers compared to the case in which piece rate remains constant.

Oppositely, the effect of an increase in piece rate that followed a decrease does not have a strong effect compare to the case in which piece rate remains constant.

Support for result 3:

We compare mean differences between the number of correct answers in period 2 minus the correct answers in period 1 for both tasks.

Mean comparison for the more challenging task (table 8) comparing the *constant p.r.* treatment with the *peak p.r.* treatment shows an estimated difference of 0.57, 95% CI [-0.07; -1.21].

Mean comparison for the less challenging task (table 9) comparing the *constant p.r.* treatment with the *peak p.r.* treatment indicates a very large effect size of 1.95, 95% CI [1.33; 2.57].

Once we compare mean difference for valley p.r. and constant p.r we observed that for both tasks the effect size reported is very small. [17]

Testing Hypothesis IV

The fourth hypothesis suggests that the effect of performance contingent monetary incentive is stronger for the less challenging task.

$$\textit{Piece rate incentive Effect (Less challenging task)}$$

$$>$$

$$\textit{Piece rate incentive Effect (More challenging task)}$$

More Challenging Task: Correct answers in period 1

Treatment	Constant p.r	Peak p.r.	Valley p.r.
Fixed Payment	-0.26	0.17	-0.39

Table 10: Testing hypothesis IV - more challenging task

Less Challenging Task: Correct answers in period 1

Treatment	Constant p.r	Peak p.r.	Valley p.r.
Fixed Payment	-2.30	-1.46	-1.42

Table 11: Testing hypothesis IV - less challenging task

Tables 10 and 11 report difference in means of number of correct answers in period 1 for the fixed payment treatment compared with each of the other treatment for the more challenging task and for the less challenging task respectively. Numbers reported in the right part of the table represent Cohen's d measuring the effect size of groups' differences.

[17] At the same time it is interesting to notice the large Cohen's d once we compare Peak p.r. treatment with Valley p.r. treatment for the less challenging task, d = -2.24, 95% CI [-2.88; -1.58]. This perhaps further indicating a detrimental effect caused by a decrease in piece rate which follows a previous increase.

Result 4:

The positive effect of performance contingent monetary incentives on effort provision is larger for the less challenging task.

Support for result 4:

We compare mean differences on the number of correct answers in period 1 for both tasks.

Table 10 compares for the more challenging task the average number of correct answers for the *fixed payment* treatment against each piece rate treatment individually.

We observe that the estimated difference between the mean of the fixed payment treatment is small compared to the other three performance contingent treatments.

Differences of the fixed payment treatment with respect to each of the other treatments are:

-0.26, 95% CI [-0.88; 0.37] compared to the constant p.r. treatment;

0.17, 95% CI [-0.47; 0.80] compared to the peak p.r. treatment;

-0.39, 95% CI [-1.03; 0.24] compared to the valley p.r. treatment;

Table 11 reports the same analysis for the less challenging task.

The estimated difference between the mean of the fixed payment treatment is much larger and in the direction predicted compared to each of the other three performance contingent treatments.

Differences of the fixed payment treatment with respect to each of the other treatments are:

-2.30, 95% CI [-3.03; -1.57] compared to the constant p.r. treatment;

-1.46, 95% CI [-2.08; -0.81] compared to the peak p.r. treatment;

-1.42, 95% CI [-2.05; -0.79] compared to the valley p.r. treatment;

6.2 Analysis with number of incorrect answers

In figure 10 and 11 we report graphically the evolution of the number of incorrect answers for each task.

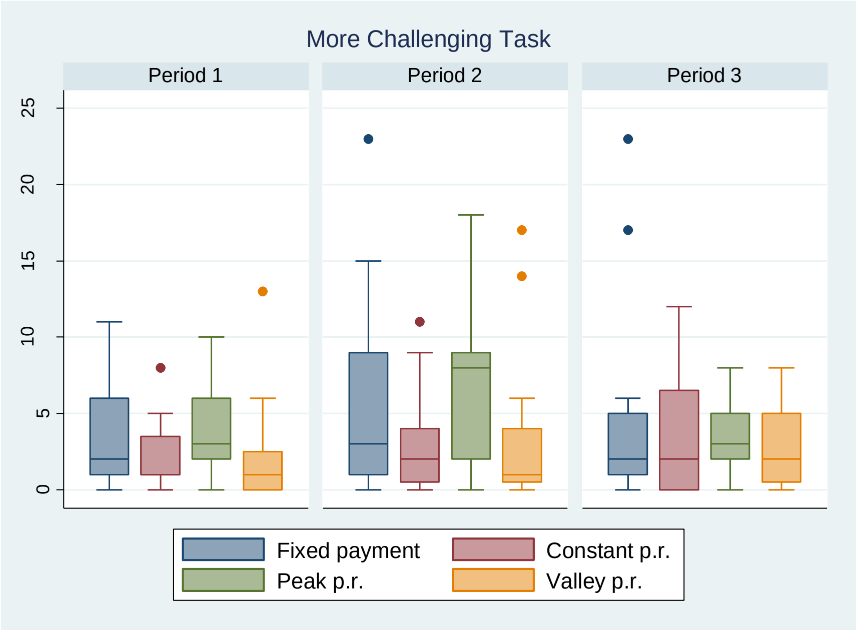


Figure 1: Number of incorrect answers for each period in each treatment - more challenging task

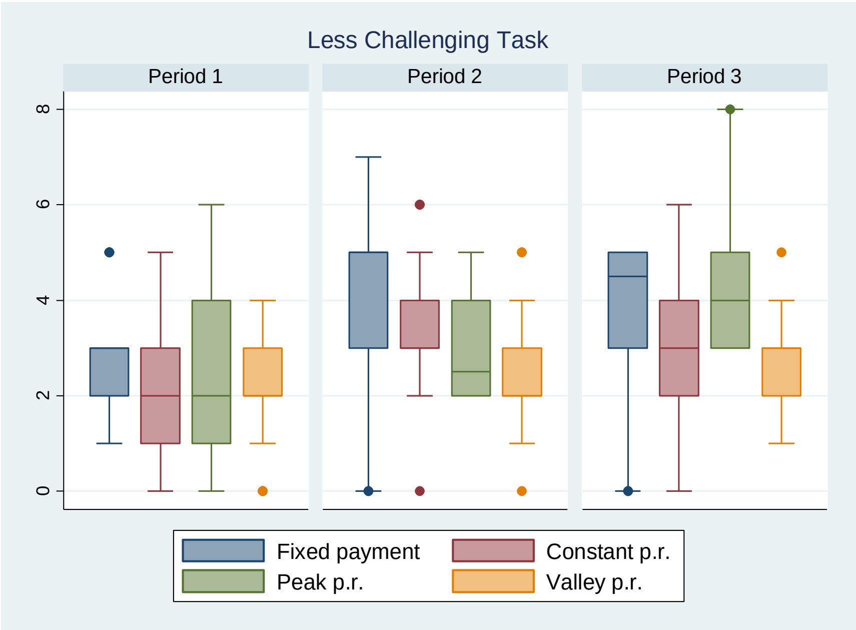


Figure 2: Number of incorrect answers for each period in each treatment - more challenging task

In figure 10 we observe the distribution of the number of incorrect answers for the more challenging task.[18]

Generally, the number of incorrect answers does not appear to be very different between treatments in period 1 and period 2. While in period 3, for the peak p.r. treatment we notice an important increase in the number of incorrect answers.

This might be the result of subjects attempting to introduce as much numbers as possible moved by the relatively higher piece rate.

In figure 11 we observe the distribution of the number of incorrect answers for the less challenging task.

A general pattern that we observe is that in late periods the number of incorrect answers increases in each treatment, probably as a result of the boredom of the task.

It is particularly interesting the way in which the number of incorrect answer increases for the peak treatment in the third period compared to the constant p.r. and to the valley p.r.

The estimated effect size for the comparison of incorrect answers given in period 3 minus incorrect answers given in period 2 for the peak p.r. against the constant p.r is quite large and equal to 1.01, 95% CI [0.46; 1.53].

The same comparison for the peak p.r. against the valley p.r. returned an estimate of 0.93, 95% CI [0.39; 1.46].

Both results confirmed that the increase the number of incorrect answers in the peak p.r. treatment compared with the constant p.r. and the valley p.r. (in the third period) was quite large.

6.3 Additional results

Post treatment task evaluation

At the end of the experiment each subject answered the Intrinsic Motivation Inventory scale (Ryan, 1982) reported in this paper in section 3.3.

For each subject we calculated the average evaluation of the task and we compare mean differences.

The estimated effect size for the difference between the more challenging task (M= 3.18; SD=1.29) and the less challenging task (M=1.68, SD= 0.87) is 1.42, 95% CI [1.23; 1.60].

[18] In section 6 we explained how to interpret a box plot.

This result highlights an important difference in evaluation between the two tasks. In particular it says that participants were on average relatively more intrinsically motivated by the more challenging task.

Gender differences

We investigate whether there are differences in ability as well as differences in response to monetary incentives between males and females.

To test for differences in ability we compare total number of correct answers given during the entire experiment between males and females across all treatments.

For the more challenging task the average of total correct was of 124 for males and 112 for females.

We find a negligible effect size for gender difference of 0.23, 95% CI [-0.21; 0.67].

For the less challenging task the average of total correct answers was of 22 for males and 24 for females with very small effect size on their differences.

In order to investigate if there were gender differences in response to monetary incentives we replicate the same analysis we carried out when testing hypothesis I while controlling for gender.

For both tasks we do not observe any gender difference.

Cost-Benefit tradeoff

One of the main objectives of this research was to inform an employer about the most efficient way for his/her firm to remunerate his/her employees with performance based types of incentive.

In table 12 and table 13 we report a cost-benefit analysis for both tasks.

We report the average of the total number of correct and incorrect answers divided by treatment. In the last column we computed the average cost incurred by the experimenter in each treatment.

More Challenging Task: Total cost for the experimenter by treatment

	Correct answers	Incorrect answers	Show up fee	Payment
Fixed payment	109	12	2.0 €	11.0 €
Constant p.r.	129	8	2.0 €	7.1 €
Peak p.r.	105	15	2.0 €	7.5 €
Valley p.r.	126	8	2.0 €	6.1 €

Table 12: Cost-Benefit tradeoff - more challenging task

Less Challenging Task: Total cost for the experimenter by treatment

	Correct answers	Incorrect answers	Show up fee	Payment
Fixed payment	14	11	2.0 €	8.0 €
Constant p.r.	27	9	2.0 €	4.5 €
Peak p.r.	26	10	2.0 €	5.5 €
Valley p.r.	23	7	2.0 €	3.1 €

Table 13: Cost-Benefit tradeoff - less challenging task

The payment column was computed by multiplying the period's correspondent piece rate (and penalty) for the number of correct (incorrect) answers while adding the show-up fee.

From table 12 we observe that among the piece rate treatments, the peak p.r. results to be the less efficient one given that it is characterized by the lowest number of correct answers and by the highest cost for the experimenter.

In the analysis of the less challenging task reported in table 13, we obtain a similar result. In particular we observe that the constant p.r. treatment provides almost the same average of correct answers than the constant p.r. but it is 20% more expensive.

7 Discussion

7.1 Summary of results

In this paper we examined the impact of different payment schemes on performance on two effortful tasks by means of a laboratory experiment.

Tasks differed since the first was designed to be relatively more challenging compared to the second. Furthermore, the first task scored significantly higher on an interest/enjoyment scale aimed at measuring subjects' intrinsic motivation toward the task.

In each task we implemented four different payment schemes separately in a three period environment. The first schema rewarded subjects with a fixed and known amount in each period that did not depend on subjects' performance.

The other three schemes were performance contingent and only differed since, in the second period, the piece rate per correct answers was doubled, halved or remained unchanged compared to period one and period three.

This simple structure allowed us to investigate subjects' response to a variety of changes in incentives.
We obtained the following results:
First we observed that in line with standard economic theory, performance contingent monetary incentives affect subjects' effort provision.
At the same time we observed that:
(1) The effect of monetary incentive is task dependent. Consistent with the literature on intrinsic motivation, we infer that monetary incentives are more effective for a less challenging task in which the level of intrinsic motivation is lower.
(2) For a unique change in incentive magnitude we found that subjects' response is stronger when the change is positive compared to when the change is negative.
(3) For two consecutive changes in incentive magnitude we observed an asymmetric response of the level of effort provided by subjects. In particular we realized that, consistent with the literature on reference point shifting, the effect of a decrease in piece rate which follows a previous increase is negative and much stronger than the effect of an increase following a decrease.

7.2 Managerial implications

Taking into account the usual concern of external validity that characterizes laboratory experiments, the study we proposed has mainly three potentially important implications for an employer looking for the best way to motivate his employees.
First, it suggests that for boring and repetitive tasks an employer should enhance employees' effort by targeting their salary on performance instead of paying them on a fixed based wage. This might allow the employer to motivate his employees more effectively.

Second, it suggests that an increase in an employee's performance based salary might positively affect employees' willingness to exert effort. Perhaps this result supports the use of positive rewards and bonuses in firms to incentivize workers.

Third, and in our view the most important suggestion, if an increase in employees' remuneration is followed by a decrease, the result can be detrimental for employees' motivation. An employer should therefore avoid substantially increasing an employee's salary if the increase cannot be sustained over time.

This result might also suggest that the effect of bonus and prizes might not have long lasting benefits.

7.3 Limits and future research

In this paper we started to investigate how different incentives' dynamics can affect subjects' willingness to exert effort.

Clearly in our experiment changes in piece rate were completely exogenous and furthermore, subjects did not receive any explanation regarding the reason for these changes.

At the same time we acknowledge that often in real life these changes come with an explanation. Think for example of a case where an employer has to cut employees' wages because he is facing financial problems.

An interesting question would therefore be to investigate if, in presence of a plausible justification, subjects will react to a change in incentive differently. More specifically, if social preferences such as altruism, reciprocity or guilt aversion could affect subjects' effort provision.

Another interesting question could be to investigate how long a damage caused by an increase followed by a decrease in the amount of piece rate lasts. This could be studied by constructing longer sequences of changes in piece rate amount in which more complete incentives' dynamics could be implemented.

References

Abeler, J., Falk, A., Götte, L., and Huffman, D. (2011), "Reference points and effort provision." *American Economic Review*, 101, 470–492.

Arkes, H., Hirshleifer, D., Jiang, D., Lim, S. (2008), "Reference point adaptation: Tests in the domain of security trading." *Organ. Behav. Human Decision Process* 105(1) 67–81

Arkes, H., Hirshleifer, D., Jiang, D., Lim, S. (2010), "A cross-cultural study of reference point adaptation: Evidence from China, Korea and the US." *Organ. Behav. Human Decision Process* 112(2) 99–111

Baker, G., Jensen, M., and Murphy, K. J. (1988), "Compensation and Incentives: Practice vs. Theory". *Journal of Finance*, 43, pp. 593-616.

Banker, Rajiv; Seok-Young Lee, and Gordon Potter. (1996). "A Field Study of the Impact of a Performance-Based Incentive Plan". *J. Acc. Econ*., 21, pp. 195-226.

Baucells, Manel, Martin Weber, & Frank Welfens (2011). Reference Point Formation and Updating. *Management Science* 57(3), 506–519

Bonner, S.E; and G. B. Sprinkle. 2002. The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Accounting, Organizations and Society* 27 (4–5): 303–345.

Bruggen, Alexander, Strobel, Martin, 2007. Real effort versus chosen effort in experiments. Economics Letters 96 (2), 232–236.

Camerer, Colin and Hogarth Robin. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor Framework," *J. Risk Uncertainty* 19, pp. 7-42

Camerer, C., Babcock, L., Loewenstein, G., Thaler, R., 1997. Labor supply of New York city cab drivers: One day at a time. Quarterly Journal of Economics 111, 408–441.

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*

Chen, H. A., A. R. Rao. 2002. “Close encounters of two kinds: False alarms and dashed hopes”. *Marketing Science*. 21(2) 178–196

Cumming, G. (2013). TheNew Statistics: Why and How. Psychological Science 25,7-29

Deci, Edward L. (1971). “Effects of Externally Mediated Rewards on Intrinsic Motivation*.” Journal of Personality and Social Psychology*, 18(1): 105–115

Deci, E., Koestner, R. and Ryan, R. (1999), "A Meta-Analytic Review of Experiments Examining the Effects of ExtrinsicR ewardso n IntrinsicM otivation" *Psychological Bulletin*, 125 (6), 627-668

Dickinson, D.L. (1999). "An Experimental Examination of Labor Supply and Work Intensities." *Journal of Labor Economics*. 17(4) Part 1,638-670.

Fehr, E., Goette, L., & Lienhard, M. (2008). Loss Aversion and Effort: Evidence from a Field Experiment. Working Paper. (University of Zurich)

Frey, B.S. (1997). On the relationship between intrinsic and extrinsic work motivation. International Journal of Industrial Organization, 15, 427-439

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-made Economic Experiments. Experimental Economics, 10 (2), 171-178

Gneezy, U. and Rustichini, A. (2000), "Pay Enough or Don't Pay at All", *Quarterly Journal of Economics*, 115(3), 791-810

Gibbons, R. and Murphy, K. (1992). "Optimal Incentive Contracts in the Presence of Career Concerns: Theory and Evidence", *Journal of Political Economy,* 100, 468-505

Holmstrom, Bengt. "Moral Hazard and Observability." *Bell J. Econ*. 10 (Spring 1979): 74-91.

Hossain, T., List, J.A. (2009). “The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations”, Management Science, vol.58 (12), 2151 -2167

Kahneman, D., A. Tversky. 1979. “Prospect Theory: An analysis of decision under risk”. *Econometrica* 47(2) 263–292

Kale, J., Reis, E., Venkateswaran, A., 2009. Rank-order tournaments and incentive alignment: the effect on firm performance. *Journal of Finance* 64(3), 1479-1512

Koszegi, B., & Rabin, M. (2006, November). A Model of Reference-Dependent Preferences. The Quarterly Journal of Economics, 121 (4), 1133-1165.

Lazear, Edward. 2000. "Performance Pay and Productivity." *American Economic Review*. December, 90:5, pp. 1346-361.

Libby, R., & Lipe, M. G. (1992). “Incentive effects and the cognitive processes involved in accounting judgments”. *Journal of Accounting Research*, 30, 249–273.

Mohd-Sanusi, Z., Mohd-Iskandar, T. (2007) “Audit judgement performance: Assessing the effect of performance incentives, effort and task complexity”. *Managerial Auditing Journal*, Vol. 22, pp. 34–52

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away from Competition? Do Men Compete Too Much?" Quarterly Journal of Economics, 122(3): 1067-1101.

Paarsch, H.J., Shearer,B.S. (2009). “The Response to Incentives and Contractual Efficiency: Evidence from a Field Experiment”. European Economic Review (53), 481-49

Prendergast, Canice. (1999). “The Provision of Incentives in Firms.” *Journal of Economic Literature*, 37(1): 7–63.

Pritchard, R. D., Jones, S. D., Roth, P. L., Stuebing, K. K., & Ekeberg, S. E. (1988). “Effects of group feedback, goal setting, and incentives on organizational productivity”. *Journal of Applied Psychology*, 73, 337–358

Pokorny, K. (2008). Pay – but do not pay too much: An experimental study on the impact of incentives. *Journal of Economic Behavior & Organization*. 66, 251-264

Rey-Biel Pedro, Roman Sheremeta and Neslihan Uler (2013).The effects of information and culture on giving. Working paper

Riedel, J. A., Nebeker, D. M., & Cooper, B. L. (1988). “The influence of financial incentives on goal choice, goal commitment, and task performance”. *Organizational Behavior and Human Decision Processes*, 42, 155–180.

Ryan, R.M. (1982). “Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*,43, 450-461

Ryan, R. M., Mims, V., & Koestner, R. (1983). “Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory”. *Journal of Personality and Social Psychology*, 45, 736–750

Shi, Lan (2007) "Productivity Effect of Piece Rate Contracts: Evidence from Two Small Field Experiments," Mimeo, Department of Economics, University of Washington, Seattle

Takahashi, H., Shen, J., Ogawa, K. (2014). “An experimental Examination of Compensation Schemes and Level of Effort in Differentiated Tasks”. Working paper

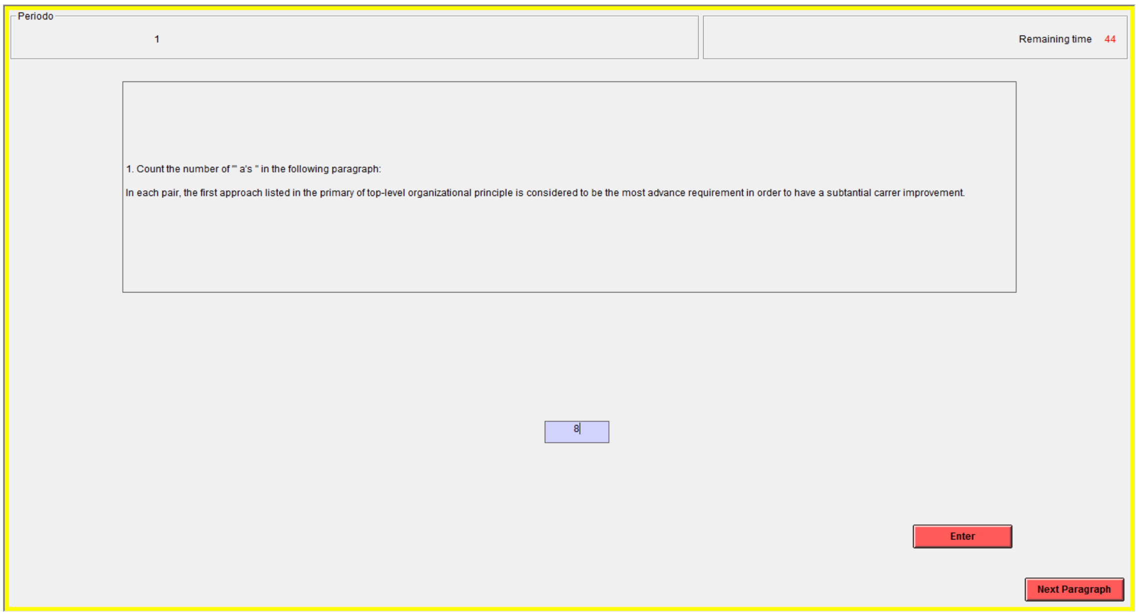
Weiner, M. J., & Mander, A. M. (1978). “The effects of reward and perception of competency upon intrinsic motivation”. *Motivation and Emotion*, 2, 67–73.

Appendix

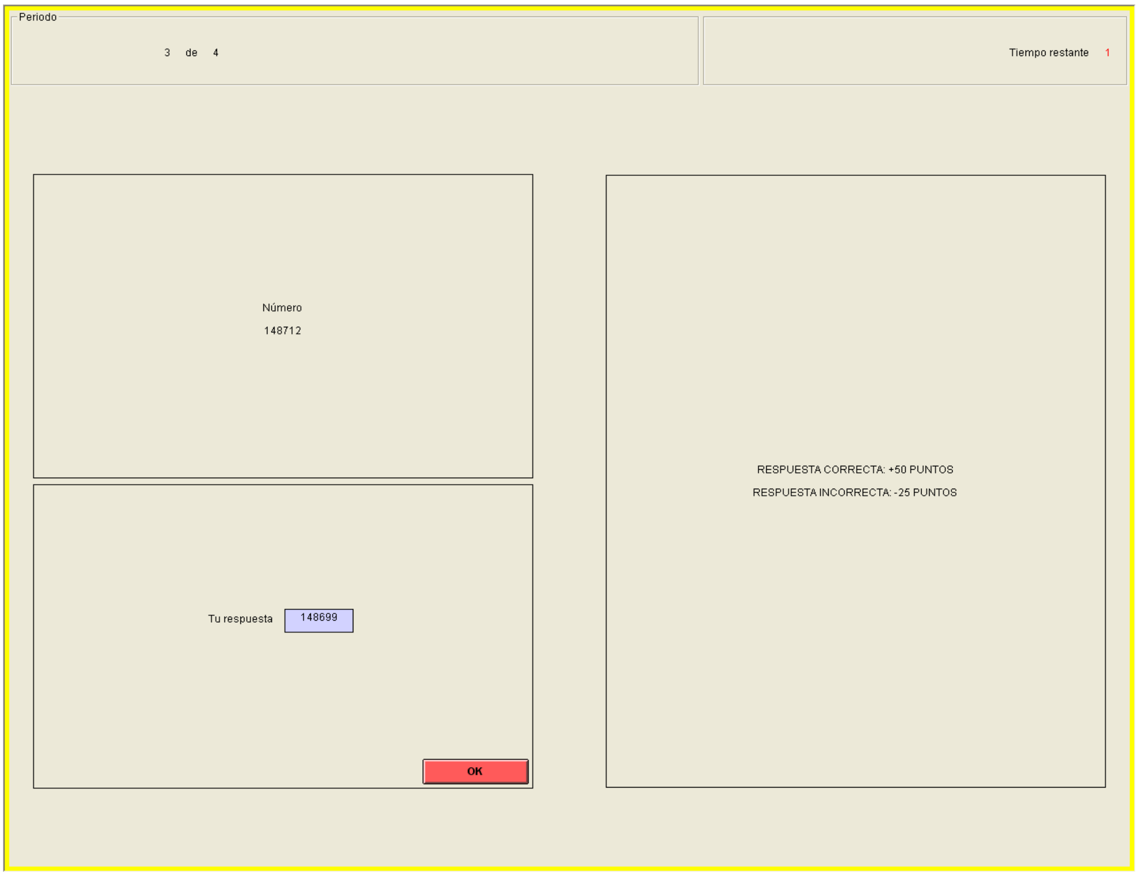
Description of the tasks.

Here we report a screenshot for the less challenging task. In this task, subjects have to count the number of A's contained in each paragraph. Once they insert the number, they have to press the button "Enter" followed by the button "Next Paragraph".

The program will sum the number of correct and incorrect answers without communicate it to the subjects.



Here we report a screenshot for the more challenging task (in Spanish). Subjects have to downward counting from the number displayed under the field "Número" each time subtracting a fixed number announced at the beginning of the period. Once inserted a number in the field "Tu respuesta" they have to press the "OK" button. The program will sum the number of correct and incorrect answers without communicate it to the subjects.


Periodo
3 de 4
Tiempo restante 1
Número
148712
Tu respuesta
148699
OK
RESPUESTA CORRECTA: +50 PUNTOS
RESPUESTA INCORRECTA: -25 PUNTOS