Hilken, K.; Rosenkranz, Stephanie; De Jaegher, Kris; Jegers, M.

Working Paper

Reference Points, Performance and Ability: A Real Effort Experiment on Framed Incentive Schemes

Discussion Papers Series, No. 13-15

Provided in Cooperation with:
Utrecht University School of Economics (U.S.E.), Utrecht University

Tjalling C. Koopmans Research Institute

Discussion Paper Series nr: 13-15

Reference Points, Performance and Ability: A Real Effort Experiment on Framed Incentive Schemes

Katharina Hilken
Stephanie Rosenkranz
Kris De Jaegher
Marc Jegers

Universiteit Utrecht
Utrecht School of Economics

Tjalling C. Koopmans Research Institute
Utrecht School of Economics
Utrecht University

Kriekenpitplein 21-22
3584 EC Utrecht
The Netherlands
telephone +31 30 253 9800
fax +31 30 253 7373
website www.koopmansinstitute.uu.nl

The Tjalling C. Koopmans Institute is the research institute and research school of Utrecht School of Economics. It was founded in 2003, and named after Professor Tjalling C. Koopmans, Dutch-born Nobel Prize laureate in economics of 1975.

In the discussion papers series the Koopmans Institute publishes results of ongoing research for early dissemination of research results, and to enhance discussion with colleagues.

Please send any comments and suggestions on the Koopmans institute, or this series to J.M.vanDort@uu.nl

ontwerp voorblad: WRIK Utrecht

How to reach the authors

Please direct all correspondence to the first author.

Katharina Hilken
Microeconomics for Profit and Non Profit Sector
Vrije Universiteit Brussel
Pleinlaan 2
1050 Brussel
Belgium.
E-mail: khilken@vub.ac.be

This paper can be downloaded at: http:// www.uu.nl/rebo/economie/discussionpapers

Utrecht School of Economics
Tjalling C. Koopmans Research Institute
Discussion Paper Series 13-15

Reference Points, Performance and Ability: A Real Effort Experiment on Framed Incentive Schemes

Katharina Hilken[a]
Stephanie Rosenkranz[b]
Kris De Jaegher[b]
Marc Jegers[a]

[a]Vrije Universiteit Brussel, Microeconomics of Profit and Non Profit Sector

[b]Utrecht University School of Economics
Utrecht University

November 2013

Abstract
The paper investigates the effect of four differently framed payment contracts on the agent's effort provision and performance in a real effort experiment. The four incentive payments are framed as a base wage and bonuses (one immediately pays bonuses, the other only after an initial performance-independent part), penalties or a combination of bonuses and penalties. The base wage that is offered, induces the reference point. The participants provide real effort and are paid for finding pairs in a customized Memory game. The bonus-only frame elicits the highest effort, whereas frames with penalties lag behind. Ability positively complements the effect of effort on performance. The combination of penalties and bonuses minimises the costs of the principal only for low levels of performance employing heterogeneous agents. For higher performance levels, framing a base wage with bonuses is cost-effective.

Keywords: Real Effort Experiment, Optimal Payment Scheme, Principal-Agent Relationship, Ability, Bonus, Penalty

JEL classification: M52, J33, C91

Acknowledgements
We thank Maarten A. Schenkeveld for the programming the experiment. We are grateful for the useful comments and suggestions of the participants of the internal seminar held at the Vrije Universiteit Brussel (VUB) and of the ESA 2013. Special thanks to the comments of Sigrid Suetens, Bruno Hendels, Mitchell van Balen and Steven De Schepper.

1 Introduction

In organisations it is common practice to use financial incentives to improve the performance of employees. As empirically shown and theoretically argued, economic agents in principal-agent frameworks indeed respond to incentives. Lazear (2000), for instance, suggests that the average level of output per worker increases between 20% and 36% after switching from fixed to variable pay, of which half is attributed to incentive effects. In a field experiment conducted in a Chinese factory, Hossain and List (2009) found that incentives significantly increase productivity in general, but more for teams than for individuals. Additionally, Fehr, Goette, and Lienhard (2008) show in an experimental setting that when piece rates are paid rather than fixed wages the output that the participants could influence one-to-one with their effort provision is about 15% higher. Similar results have been found by Shearer (2004).

By introducing prospect theory, Kahneman and Tversky (1979) suggested several behavioural biases that have gained increasing importance in contract theory and its applications: rather than being concerned with absolute outcome levels, agents evaluate outcomes relative to a reference point as gains or losses. Moreover, losses hurt the decision-maker more than gains give pleasure (the concept of loss aversion) and decision-makers are predominantly risk averse in the domain of gains, but become risk seeking in the domain of losses, a phenomenon known as the reflection effect. This effect is in line with diminishing sensitivity, which suggests that the difference in utility obtained from any two outcomes diminishes with the distance to the reference point. There is increasing experimental evidence that reference points and framing indeed influence actual behaviour (see Kühberger, 1998 for a detailed survey on the framing literature). Strategically influencing the reference point could then lead to improvements in the productivity of the agent. The understanding of the implications of these findings and the use of strategic framing is still limited. We discuss more recent literature on framing in both theoretical and experimental research in the following paragraph.

When designing incentive schemes to increase the employee's effort the question arises how to incorporate the above-described concepts: what kind of incentive scheme should be used? What determines an agent's reference point when confronted with an employment contract? Do bonuses work better than penalties, or should both be combined? Which *frame* is more effective? Several theoretical and experimental studies have attempted to answer these questions: Herweg, Müller, and Weinschenk (2008) develop a model that predicts binary payment schemes with a base wage and bonuses to be optimal when, following Kőszegi and Rabin (2006), the agent's stochastic reference point is determined by his expectations. Abeler, Falk, Götte, and Huffman (2009) influence the expectations of their participants and find that if expectations are high, agents provide higher effort compared to low expectations. Armantier and Boly (2012) suggest in their experimental study that framing payments as (small) penalties increases effort only when combined with bonuses. The majority of experimental studies, however, suggests that using penalties in payment schemes increases effort: Hannan, Hoffman, and Moser (2005) for instance research the difference between economically equivalent bonus and penalty contracts and find that there are two mechanisms via which framing affects employee effort: One the one hand, the higher the expected disappointment with not receiving a bonus or having to pay a penalty, the lower employee effort, mediating the effect of the contract frame. On the other hand, perceived fairness of the offered contract has a positive effect on employee effort. The first effect, based on the notion of loss aversion, dominates the second effect of reciprocity, which results in higher employee effort under the penalty scheme. Similarly, Fehr, Klein, and Schmidt (2007) highlight the importance of fairness in moral hazard settings. Hannan et al. (2005) in the laboratory, as well as Hossain and List (2009) in a field experiment, compare a bonus and a penalty contract, but do not consider combinations of these reward mechanisms. Fehr and Schmidt (2007) suggest that combining penalties with bonuses might be less efficient than pure bonuses. Hossain and List (2009) find that framing binary payments as penalties increases productivity for both individuals and teams compared to framing payments as bonuses. Equivalent results have been found by Fryer, Levitt, List, and Sadoff (2012) when framing is applied to teachers' incentives. Non-participation is not possible in the majority of these experiments as either a non-zero effort is to be chosen or because the employees already have a contract in the administered field experiments. Based on the theoretical predictions of Hilken, De Jaegher, and Jegers (2013), bonus incentives are optimal because before penalty incentive effects can work in favour of the principal, the agent will turn down the contract.

This paper is set out to analyse whether framing payments as bonuses rather than as penalties elicits higher effort by the agent (which then translates to higher performance): making the agent feel losses with framing penalties should intuitively lead to higher effort provision because of the presence of loss aversion in the incentive constraint. Yet, if the reservation utility of the agent is not affected by the reference point, then the more penalties are perceived, the higher the payment that the principal needs to offer to make the agent participate. This participation effect is so strong that it always overcompensates any incentive effect that could come about from framing penalties. These theoretical predictions are supported by the experiment conducted by Luft (1994) on contract preferences which showed that agents are more likely to participate in bonus contracts rather than in equivalent penalty contracts. For the purpose of researching these predictions, we designed a real effort experiment in a between-subject design. We comprehensively assess four differently framed payment schemes with the same base wage as to their effect on effort: 1) a frame with base wage and bonuses for all performance levels, 2) a frame with base wage and bonuses for higher performance levels, 3) a frame with penalties for lower performance levels and a base wage for higher levels and 4) a frame with penalties for low performance levels, base wage and bonuses for higher performance levels. We find that the frames have a significant effect near the the base wage (the reference point), but that further away from the reference point, their effect vanishes. On these ranges where a treatment effect is present, framing payments as bonuses only is optimal for the principal in inducing higher effort, framing payments as penalties results in the lowest effort. In contrast to earlier studies, we find framing payments as bonuses is most effective to induce higher effort by the agent. Effort in turn has the greatest positive effect on performance. We furthermore shed light on the importance of task-relevant ability, as Bonner and Sprinkle (2002) call for and which distinguishes our research from the existing literature: the ability of an agent has a positive effect on performance, by complementing effort. The higher the ability of an agent, the greater the effect of the exerted effort on performance.

The paper is structured as follows. In Section 2, we describe the theoretical background and the hypotheses. We transpose the theoretical predictions to the real effort experiment whose design is described in Section 3. Results are presented in Section 4 along the lines of the hypotheses. In Section 5 we discuss our results and address the effects of other variables identified in the previous analysis. Section 6 concludes.

2 Predictions

We base our experimental design in Section 3 on the theory of De Meza and Webb (2007), who model reference-dependent preferences in a principal-agent relationship. They derive four incentive payment schemes that are equivalently optimal with respect to performance.[1] Strategic framing, the intentional manipulation of the reference point, can alter the outcomes significantly (see the Asian Disease experiment, Tversky & Kahneman, 1981). Additionally, Hilken et al. (2013) investigate the effect of strategic framing in a principal-agent relationship. We shortly describe the two models and derive our main hypotheses.

De Meza and Webb (2007) determine the optimal incentive schemes by minimising the principal's costs subject to the participation constraint and the incentive compatibility constraint of the agent. The difference to the standard solution of Holmstrom (1979) is that the authors introduce a loss utility function in addition to the standard utility derived from the realised income. The realised income is compared to the reference income by the agent. If the realised income lies below the reference income, the agent feels a loss and his utility is diminished. In their first

[1] The authors discuss three formulations of the reference income. We focus only on the first proposition with an exogenous reference income.

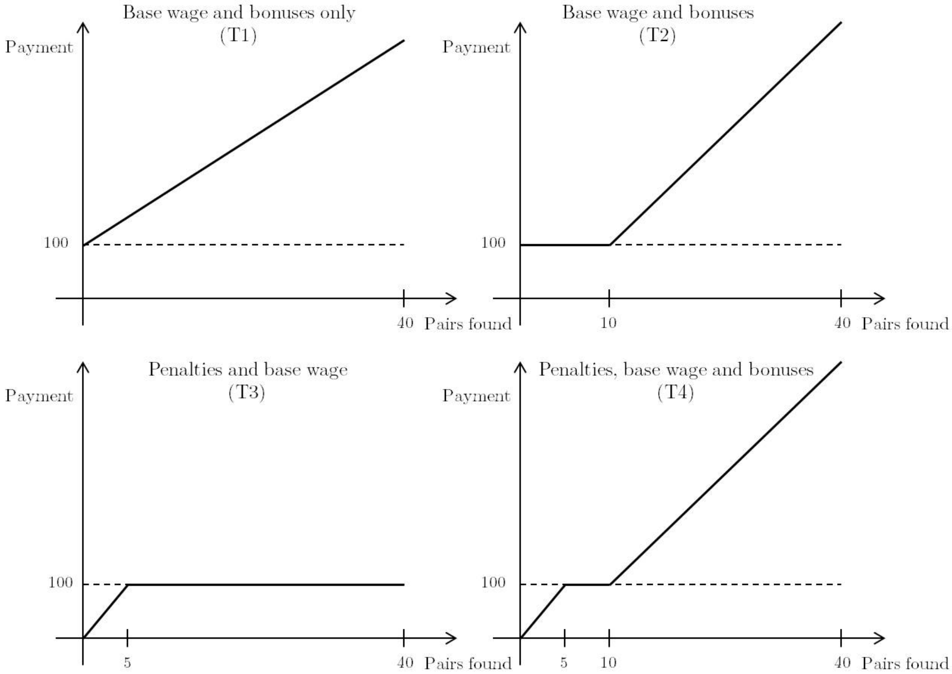


Figure 1: The four payment schemes.

formulation of the problem, De Meza and Webb (2007) assume linear loss aversion and an exogenously determined reference point. The overall utility of the agent in the loss region is therefore: $U(Y) - [l(U(Y^R) - U(Y))] - c(e)$. Y is the realised income, Y^R is the reference income, $l > 0$ is the loss aversion measure and $c(e)$ is the cost of effort. In the gain region utility does not contain the second term in square brackets. The agent has the choice between two effort levels (high and low), $e \in \{\underline{e}, \bar{e}\}$. With high (low) effort, the chance of achieving higher outcomes, $f(s|e)$ with $s \in [\underline{s}, \bar{s}]$, for the principal is larger (smaller).[2] Deriving the first-order conditions and the solution to the principal's optimisation problem results in four possible optimal incentive payment schemes. De Meza and Webb (2007) characterise these optimal payment schemes in their Proposition 1 (p.73). The first payment schemes pays the reference wage and bonuses for increasing performance. The second payment scheme pays the reference wage up to some cutoff point and bonuses for higher performance levels. The reference wage is paid for higher performance levels and penalties are deducted for low performance levels in the third payment scheme. Finally, the fourth payment scheme combines bonuses and penalties: between two cutoff points the reference wage is paid from which penalties are deducted for lower performance levels and bonuses are added for higher performance levels. These four payment schemes are depicted in Figure 1 with the thresholds and payments to the agents that are used in our experiment. Because of loss aversion, the kink in the utility function of the agent at the reference point causes the payment schemes to be performance-invariant at the reference point. This in turn makes designing economically equivalent payment schemes impossible. We address this problem with changing the exchange rate across treatments.

In Hilken et al. (2013), an additional first-order condition enters the minimisation problem of the principal to reflect the effect of strategically framing the reference point. They assume that the reference point of the agent can be influenced by the principal by appropriately choosing the wording of the payment contract, specifically in terms of the combination of the base wage and bonuses and/or penalties. In this case the base wage, which can be influenced directly by the

[2]The Inada condition and the monotone likelihood ratio property are assumed to hold. The principal is risk neutral and the agent is risk averse. The reservation utility enters the participation constraint of the agent.

principal, determines the reference point of the agent.[3] Hilken et al.'s (2013) results show that of the four incentive payment schemes derived by De Meza and Webb (2007), with strategic framing only the strictly increasing one is optimal. With strategic framing, the principal is able to directly change the reference point of the agent by setting the base wage. In consequence, the agent's decision to increase or to decrease the reference point also influences the way he/she perceives the performance-dependent payments. By increasing the reference point, the principal automatically moves payments into the loss region as defined by the solution to the optimisation problem. When giving penalties to the agent, the agent becomes loss averse, which is costly to the principal. To compensate for the feeling of loss for some low performance level payment, the principal needs to increase one of the payments in the gains region *by more* than the penalty. This means that the principal should set a low base wage and give bonuses for all possible performance levels in order to induce an optimal effort level. Based on the predictions of these models, we formulate our first hypothesis:

Hypothesis 1. *When payments are strategically framed as gains, performance is higher than under all other payment schemes.*

The question we aim to answer with this experiment is whether any frame induces the agent to provide higher levels of effort than other frames. A factor that may affect the optimality of incentive schemes, but which is not modelled by neither De Meza and Webb (2007) nor Hilken et al. (2013), is the agent's ability. Firstly, for a given level of effort, the high-ability agent (subscript H) has a greater probability of success compared to the low-ability agent (subscript L), e.g. $f(\bar{s}|e)_H > f(\bar{s}|e)_L$. Therefore, with the same level of effort, a high-ability agent may reach higher levels of performance, or put differently, the same level of performance may be reached with lower levels of effort, independent of the payment scheme employed. This reasoning leads to the following two (interrelated) hypotheses:

Hypothesis 2.

a, *Higher-ability agents show higher performance levels at the same level of effort when compared to low-ability agents.*

b, *Higher-ability agents show lower effort levels at the same level of performance when compared to low-ability agents.*

Secondly, with higher ability, the costs of effort are lower for every level of effort (depicted in Figure 2). The assumption that the difference of the costs of effort for given levels of effort is lower for the high-ability agent than for the low-ability agent implies that the former group has lower marginal effort costs, e.g. $c(\bar{e}_L) - c(\underline{e}_L) > c(\bar{e}_H) - c(\underline{e}_H)$. As an agent chooses the optimal effort level such that marginal costs are equal to marginal revenues, bonuses increase effort for all agents (as compared to a payment scheme without bonuses) because of their positive effect on the marginal revenue. However, this effect should be greater for high-ability agents. In contrast, penalties have a negative effect on the marginal revenue and therefore decrease effort for all agents, and this effect should be greater for low-ability agents for the same reasons. The combination of these two effects could lead to a cancellation when bonuses and penalties are combined. This leads to the following hypothesis:

Hypothesis 3. *Payment schemes that frame*

a, *gains induce higher performance of high-ability agents,*

b, *losses induce lower performance of low-ability agents,*

[3]Although not explicitly mentioned, this assumption is made by most framing experiments, such as Hannan et al. (2005), Armantier and Boly (2012) or Fehr and Schmidt (2007). The participants receive some amount at the beginning of the experiment, which can be interpreted as a base wage, and bonuses or penalties are framed according to this amount. This amount is then interpreted as the reference point. At the end of the experiment, bonuses or penalties are added/deducted according to the participants' performance.

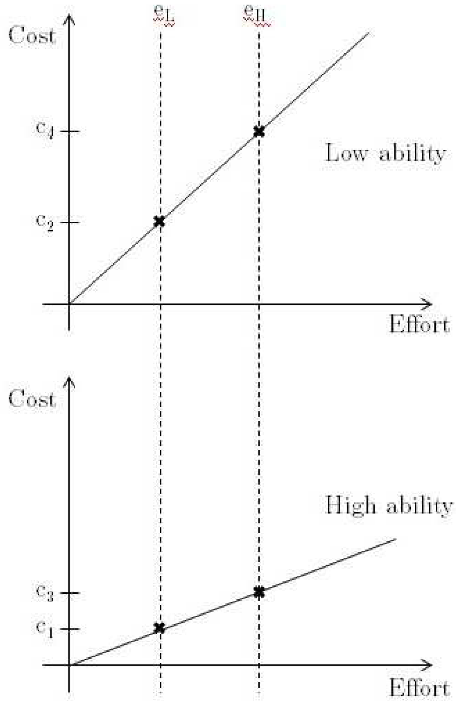


Figure 2: Cost functions for different ability levels

c, gains and losses combined induce similar performance of high- and low-ability agents.

The objective of the principal is obviously to maximise profits. Suppose now that there is a single output level and capital is fixed in the short-run. The profit maximisation problem of the principal is then equivalent to minimising the cost of labour (wage times labour, $w * L$), subject to the constraint that the labour input can at least produce a specified output level (Q):

$$min_{L \geq 0} \qquad w * L \tag{1}$$

s.t.

$$F(L) \geq Q$$

When the output required (Q) is the maximum capacity, the principal chooses the least costly combination of wage and labour for this output level. We expect that if low cumulative output levels are required only, framing payments as a high base wage and penalties might be cheaper to the principal than framing them as a low base wage and bonuses. The intuition behind this is that the agents dislike penalties because of loss aversion and consequently they make sure with the highest possible effort level to reach the performance level that at least pays the base wage. After that though the agents are expected to stop immediately. Similarly, if high cumulative performance levels are required, we expect that framing payments as bonuses with a low base wage is motivating and encouraging high effort more than framing payments as using penalties. Therefore a bonus scheme should be more efficient with respect to the costs of the principal than a penalty scheme. This rationale is formulated in the following hypothesis:

Hypothesis 4. *For low required cumulative performance levels, penalty contracts minimise the costs of the principal when employing heterogenous agents while for high required cumulative performance levels, bonus contracts minimise the principal's costs.*

We expect that different levels of risk aversion and loss aversion have an impact on how frames affect the participants' performance, but both these variables will not change the general results meaningfully. We will explain these factors in more detail in the discussion section.

3 Design

The experiment has five parts, which will be described in more detail in the following subsections. Firstly, we test the participants' ability to remember with a visuo-spatial memory test. Secondly, a Memory game is played for four rounds. In the first two rounds payments are fixed (performance independent, to induce a reference point). In the last two rounds payments are variable (depending on performance). As a treatment variable, we test four different variable payment schemes as to the effort provision and the performance of the participants. The Memory game is followed by two lottery decisions in the third and fourth part, one to assess participants' risk aversion (according to Holt & Laury, 2002), the other to assess their loss aversion (with a modified version of Gächter, Johnson, & Herrmann, 2007). The experiment ends with a questionnaire on demographics, perceived difficulty and pleasure. The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007). ORSEE was used to recruit participants. The experiment was conducted in English.

3.1 The Visuo-Spatial Memory Test

To test the short-term memory span of the participants we designed an incentivised task for the participants that requires the storage of short-term visual and spatial information.[4] From this test we derive our measure for ability independently from the following real effort task. The first screen shows the participants 9 grey squares (3 by 3) of which 3 randomly turn red for 1 second. The task of the participants is to reproduce the pattern by clicking on the previously red squares. The number of squares to be found (for example 3 red squares) in each round equals the total number of trials (e.g. 3 clicks for this example), which does not allow participants to correct themselves. They do not receive immediate feedback on their performance. In a total of 10 rounds, we increase the number of grey and red squares from round to round up to 56 grey squares of which 12 red ones have to be recognised. There is a short break of 3 seconds between the rounds. We record the number of correctly identified squares in each round and the round number. This visuo-spatial test requires the participants to remember the relative position of the squares in the two-dimensional space, or the pattern. The information only needs to be stored for a short period and therefore does not enter the long-term memory. The following Memory game exactly requires the participants to activate their short-term visuo-spatial memory: they need to link the position of an image to a position on the screen and recall the position when the same picture is found in another location. The participants receive up to 10 points for each correct square identified.[5]

We position the test at the beginning of the experiment to overcome that the participants are tired and that they do not have any incentive to complete these rounds. We do not give feedback on their performance in the visuo-spatial span test until the end of the experiment, as we do not

[4]This test is a variant of the dot memory task (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001). The original tests show a 5 times 5 grid with two to seven dots that appear for 750 to 850 ms. The participants then have to indicate the location of the dots after they have disappeared. This test does not require explicit processing of information. We modified this test to one that changes in size. Miyake et al. (2001) tested 167 participants which had to complete a total of 30 trials with a maximum of 135 correctly identified dots; mean=87.36, standard deviation=13.28, range=47-121, skewness=-0.51, kurtosis=0.50 and reliability=0.83

[5]An alternative test to the dot memory task is the Corsi block-tapping test (see Berch, Krikorian, & Huha, 1998): the sequence in the Corsi block-tapping test is predetermined and the same for each tested person. In our experiment we randomise the blocks that turn red. Self-correction is not allowed in this version of the test, contrary to the Corsi block-tapping test. Our test has the advantage that the participants only focus on the task on the screen and are not influenced by the experimenter or the environment as much. In addition to this, the time that the blocks are shown is exactly the same with the computerisation of the test. There is also no communication between the experimenter and the participant that could distort the short-term memory capacity of the participants.

want it to have any influence on the ensuing Memory test. The Memory game (described in detail in Section 3.2) could also be an indicator for ability, but it has the serious disadvantage that performance can be reached by trial and error and therefore the success is not only dependent on the participants' visuo-spatial ability.

As the rounds are not *directly* comparable as to their difficulty due to the change in size and the number of red squares in each round, we derive a single ability score for each participant in the following way: we estimate the Kernel density for each of the possible red squares to be found on the basis of all participants and we do this for each of the ten rounds in the ability test. From the Kernel density we derive the probability of finding each possible number of red squares. The difficulty of finding, for example, 12 red squares can then be determined as the probability of finding maximally 11 squares.[6] We multiply these cumulative probabilities with the maximum number of red squares in each round. We sum these numbers for each participant's performance to get an individual measure of the ability to remember, that we use in the subsequent analysis. This score is a more sensitive measure for ability than just the number of rounds they managed to complete correctly. Finding one out of three red squares in a grid of 9 squares is easier than finding one out of 12 red squares in a grid of 56 squares. The difficulty though does not necessarily double from round to round and for each number of red squares, meaning that the 10th round is not necessarily ten times harder than the first round. With the Kernel density estimation we try to get an as exact as possible scaling of the difficulty of reaching the observed performances from round to round. This procedure makes the rounds comparable to each other. An example on how the ability variable is calculated, based on the last round, can be found in appendix 7.5.

3.2 The Memory Game

The experiment is designed to research the effect of setting a base wage (equivalent to the reference income) and bonuses and/or penalties on effort provision of the agents and the resulting performance. Different verbal formulations of the payment schemes (strategic framing)[7] in each treatment with the same base wage are expected to induce different effort provisions and subsequently performance levels of the agent. The question to be answered is which payment scheme induces the agent to provide which level of effort and performance. The higher the performance of the agent, the higher the profits of the principal. Translated to the real effort task, this means that the harder the agent tries (effort), the more often he will succeed (performance). This in turn increases the revenues of the principal by increasing the output (in this case by participants finding pairs).

The *Memory* game is a deck of two times 20 pictures with dots turned upside down. By flipping two cards at a time, the participants' task is to find matching pairs. On the base of the number of found pairs, the participants receive their payment. The participants are allowed to turn cards up until they have found two pairs, after which the cards are mixed again and the Memory game starts again. We do not let the participants play longer than the first two pairs found, as the effort cost (and thus cost of finding pairs) decreases with the number of found pairs. If we would let the participants continue after two found pairs, they would possibly play the game until the end and we could not measure performance. A button on the screen gives the participants the possibility to stop whenever they want. The number of trials and the number of found pairs are our variables for effort provision and performance, respectively. Effort includes the participants' trials that do not lead to success, e.g. the trials before clicking the stop button and after the last pair found.

After one trial round without payment, we ask the participants to play the customised version of the Memory game in total 4 times with different payment schemes. The first two times we offer a fixed payment, e.g. the payment is independent of either effort or performance of the participant. These first rounds are designed to induce a reference point by giving a fixed wage of 100 points. Subsequently, for the following two rounds the participants receive a variable payment, dependent on their performance. This variable payment is expressed in the same manner, first

[6] This is equivalent to one minus the probability of finding 12 red squares or more. The probability of finding at least 12 red squares can be interpreted as the ease of finding 12 red squares.

[7] The actual formulations of the different payment schemes can be found in the Instructions, Appendix 7.1.

naming the base wage and subsequently the variable part. In total, four payment schemes are tested and compared in a between subject design. More rounds in each treatment would confuse the participants as the pictures used in the different rounds are the same. Furthermore we think that the game should be tedious (reason why we do not let them play only once), but should not make them annoyed. The information concerning the payment scheme is given separately for each treatment in an envelope on coloured cards. The exchange rate is varied in the different treatments to ensure that the participants across treatments receive approximately the same total payment. We had either 39 or 40 participants playing with one of the variable payment schemes. For an overview of the sessions run in the laboratory, see Table 9 in the Appendix.

The two rounds with a fixed payment serve as a control to see how much the participants liked the Memory game, as the theoretically expected effort and performance of the participants is zero. We can infer from these rounds whether and how much the participants are intrinsically motivated. The effort and performance should be paralleled with the answer given later in the questionnaire as to the pleasure level the participants feel while working. Although standard theory predicts that participants should not provide any effort (and should therefore not find any pairs) with a fixed payment, we expect that the participant feels pleasure with his work and/or feels bad with receiving money for not providing effort. In consequence to this reasoning, in the fixed payment rounds, the participants will (try to) find pairs irrespective of the payment.

We test the difference in performance of the participants in four different payment schemes. In the first treatment, the participants get a *contract with a base wage and bonuses only (T1)*: For each pair found, participants receive 20 points additional to the base wage of 100 points. In the second treatment participants are paid the base wage for the first 10 pairs and a bonus of 20 points for each pair above 10 (*contract with base wage and bonuses, T2*). Treatment three is the mirror image of the previously described payment scheme: 20 points per pair are deducted from the base wage of 100 points for performances below 5 pairs. Above 5 pairs, the payment scheme is insensitive to performance and pays the base wage of 100 points (*penalties and base wage contract, T3*). The last scheme is a combination of the bonus and penalty contract. The participant receives an insensitive pay for the range of pairs 5 to 10 pairs, below and above these thresholds the payments are increasing in the number of the pairs found, by 20 points (*contract with penalties, base wage and bonuses, T4*). The payment schemes are depicted in Figure 1.

Before the participants are allowed to start each treatment, they are asked two questions as to their understanding of the payment scheme. If the answer to the questions are wrong, they need to re-read the instructions on the payment cards and enter the correct answers. It is crucial for the experiment that the participants understand and internalise the payment scheme. We gave the participants cards to give them the opportunity to be able to (repeatedly) look at the payments they receive for their performance and make fully informed decisions. The participants can proceed at their own pace. When they finish, they receive their payments according to their performance. The participants do not receive any information with respect to their payments after each round as we do not want to have any influence on the induced reference income. The only performance information that the participants receive on the screen are the number of trials and the number of pairs they have found. With this we want to make sure that the participants take their effort provision decision fully informed and in a *rational* manner. Between the two variable payment rounds we asked the participants about their expectations of the earnings in the next round. As there is substantial disagreement in the literature as to the nature of the reference point, we want to make sure that the base wage is accepted as the reference point in the experiments. The participants should have no expectations about the earnings in the experiment before they start. Therefore we assume the base wage to be a good approximation of the reference point. We critically look at this assumption in the discussion section.

Performance in the Memory game is reflected by the number of pairs found and the effort by the number of trials. Therefore, we did not add an artificial cost function for the participants. Part of the performance is dependent on chance as the program randomly shuffles the cards. It might therefore happen that a participant turns two identical cards by sheer luck. On the other hand, the participants will also most likely hit on a card where they mistakenly thought the matching one was lying. Therefore, the cards are shuffled after two found pairs. We would only have a

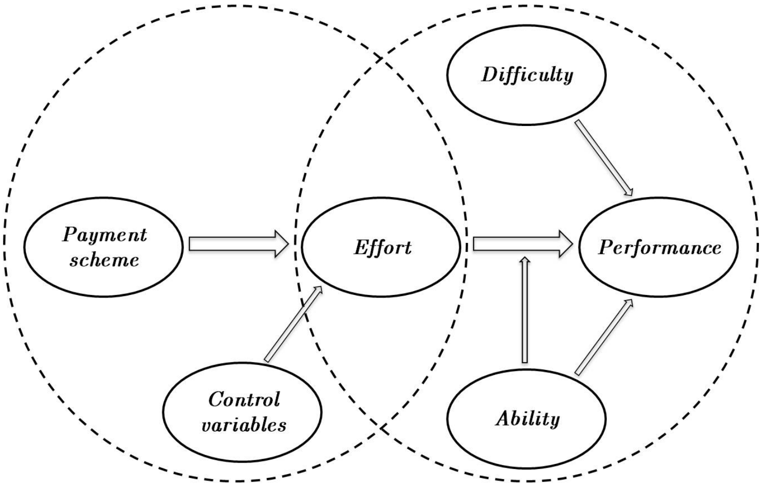


Figure 3: The Model

constant cost function if we would shuffle them after each pair found. But then the problem is that the performance becomes less dependent on the ability tested in the short-term memory task. We therefore increased the number of pairs found to two before the Memory game starts again. A round ends when the participant takes the decision to stop by clicking on a button. We collect the number of trials until one pair has been found, for each pair and each participant.

3.3 Risk aversion and loss aversion

To assess participants' risk aversion and loss aversion we administered two sequences of lottery decisions. The first asks the participants to choose between two lotteries in ten different cases (the lottery is equivalent to the one in Holt & Laury, 2002). The lotteries and the subsequent risk aversion measures are shown in Appendix 7.6 as well as the categorisation in Table16. The amounts do not change in the ten cases, only the probability of occurrence of each amount changes. The variance of the first lottery is smaller than of the second. The switching point from one lottery to the other reveals the individual risk aversion. One of the ten lotteries is chosen randomly by the computer and subsequently the lottery is played to determine the payoff for the participant in this round.

The second lottery measuring the loss aversion of participants is based on Gächter et al. (2007), where we change the absolute amounts that the participants can earn in the lottery, the expected payment though is the same. The participants have to decide on either rejecting or accepting a lottery, for six different lotteries. The computer again chooses one of these six lotteries. If the participant accepted the lottery, a virtual coin is tossed to determine the payoff of the participants. If the participant rejected a lottery, the payment is automatically equal to zero. The categorisation of the outcomes is shown in Table 18.

4 Results

In this section we statistically assess the model schematically depicted in Figure 3. In most experiments, effort cannot be observed (or at least not differentiated from performance). Therefore, these statistical analyses focus on the effect of the payment scheme on performance, although the effect of the payment scheme on performance works *via* effort (see Bonner and Sprinkle (2002) for a similar, but broader, conceptual framework). For this reason (and because we are able to measure effort in our main task), we look at the effect of the payment scheme on effort in a first regression and subsequently at the effect of effort on performance in a second regression. The demographic variables, risk aversion, loss aversion and several treatment-specific variables serve as controls in the first regression. We suppose that ability and the self-reported difficulty of the task influence how effective the effort is in achieving any performance level which are therefore included in the second regression. We discuss the effect of ability in Section 4.3.

4.1 Sample description

In total, 10 sessions were run between December 2011 and March 2012. Participants were students of various fields of study from Utrecht University. The students were not allowed to participate more than once. A total number of 157 students participated in the experiment, on average 16 per session.[8] In each session only one treatment (variable payment scheme) was administered, with two sessions per treatment. Two additional sessions were run with different payment schemes to compensate for the no-shows in previous sessions. A complete overview of the different fields of study and the payment schemes in each session is given in Appendix 7.3. The experiment lasted from 1 to 2 hours, depending on the decisions taken by the participants, who could leave the experimental laboratory whenever they liked. They received 11.79 Euros on average, with a minimum of 3.92 Euros and a maximum of 25.75 Euros. We use non-parametric statistics as our variables are either categorical or not normally distributed.

Table 1: Descriptive statistics

N=157	Mean	SD	Min	Max	1	2†	3	4	5	6
1 Age	23.17	3.72	17	42	1.00					
2 Male†	0.45	0.50	0	1	-0.06	-				
3 Pleasure	5.63	2.39	1	10	-0.01	-0.31	1.00			
4 Risk aversion	5.66	2.09	0	10	-0.08	0.45**	-0.01	1.00		
5 Loss aversion	3.29	1.08	0	6	-0.03	0.30*	0.07	0.11	1.00	
6 Total profits	11.79	4.35	3.92	25.75	0.09	-0.15	0.02	-0.06	0.13	1.00

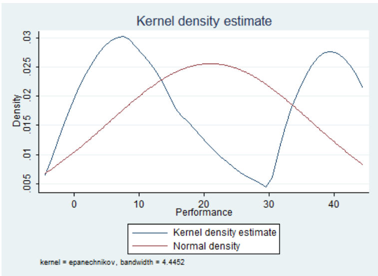
Notes: Significance levels: *p<0.10, **p<0.05, ***p<0.01
Spearman's ρ for all statistics, except †Mann-Whitney U test and Cohen's d for gender differences

The descriptive statistics and pairwise rank correlations in Table 1 give first indications for relations between the variables collected: the 157 participants are on average 23 years old and 45% are male. On the scale from 1 to 10, the participants indicated that their mean pleasure in the main task of the experiment was 5.63. More than half of the participants are Dutch (53.5%). We include these variables in the following regression analysis as control variables. The majority of the participants (53.5%) is risk averse, only 7.64% are risk loving, 23.57% are risk neutral. A high share of 15.29% is highly risk averse. In the loss aversion lotteries, 42.02% reject the fair lottery and 91.72% in total accept only positive expected value lotteries. One participant accepted all lotteries and one participant rejected all.[9] From Cohen's d we can deduct that males are significantly less risk averse (standard finding, see Croson & Gneezy, 2009, Holt & Laury, 2002 or Borghans, Heckman, Golsteyn, & Meijers, 2009) and loss averse than females, which does not seem to translate to gender effects in total profits though. We include risk aversion as well as loss aversion as categorical variables in the regression analysis as we are interested in qualitative rather than quantitative results with respect to them. We address these in the discussion section of this paper.

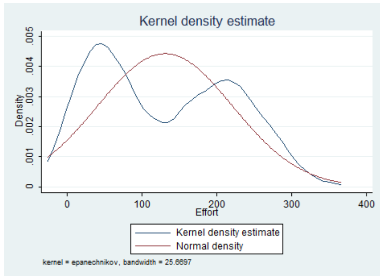
The median individually perceived difficulty had a value of 7 on the scale from 1 to 10. We reject the null hypothesis of normality with the Shapiro-Wilk test (W=0.96, z=5.19, p<0.01; also see Figure 4dd,). The construction of the ability variable from the data gives a measure that ranges from 3.06 to 42.47 points with a median of 22.03 points. From Figure 4cc, we see that the Kernel density estimate of ability has a dent around the mean of the normal distribution. The null hypothesis of ability being normally distributed is rejected (Shapiro-Wilk test, W=0.98, z=1.98, p<0.05). The ability measure is negatively correlated to the perceived difficulty of the main task (see Table 2). As the short-term memory test is designed to measure the ability of

[8] A first trial session has been run that we exclude from the analysis because of changed design.
[9] The classification from the 10 choices into the four categories is shown in Appendix 7 as well as the percentage choices for each level of risk aversion and loss aversion.

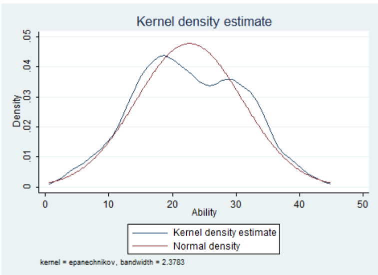
participants, a measure independent of the participants' own perceptions, this correlation shows that they can (partly) assess it themselves. It shows that with increasing ability, the main task becomes easier (in their own perception). This is in turn reflected in the profits the participants earn: ability is also positively correlated to the profits earned in the main task, which gives a first hint at the causal relationship we expected (Spearman's ρ=0.23, p<0.01). Difficulty is negatively correlated to the effort of the participants. This reflects that effort costs are increasing and higher for participants who perceive the task to be difficult. This is partly offset by ability, suggesting that more able participants face lower effort costs.



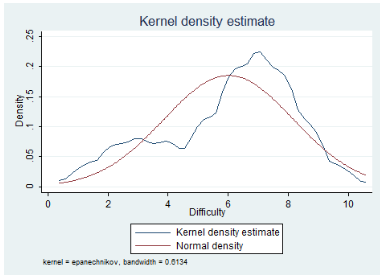
(a) a, Performance



(b) b, Effort



(c) c, Ability



(d) d, Difficulty

Figure 4: Kernel density and normal distribution plots

The distributions of our main (dependent) variables are *not* normal, which gives a first indication of non-normality of the error term in an OLS-regression.[10] Therefore, we use non-parametric regressions in the evaluation of the hypotheses: performance and effort are bimodal as can be observed in the Kernel density plots in Figures 4aa, and 4bb,. The Shapiro-Wilk test for normality confirms this visual impression (W=0.92, z=6.88, p<0.01 and W=0.93, z=6.49, p<0.01, respectively). The observed performance variable ranges from 0 to 40 which is partly predetermined by the maximum performance level per round set at 40 pairs. Effort has no higher bound as the participants have the possibility to stop the task at any time. We observe a maximum effort of 341 trials with a median of 122.5 trials. Effort and performance are highly, and significantly, correlated as can be seen in Table 2. This is not surprising as flipping cards in the real effort task should eventually lead to finding a number of pairs. Nevertheless, we suppose that a participant's ability and the perceived difficulty of the task have an effect on his/her productivity. The significant (at the 1% level) rank correlation coefficients in Table 2 between performance and ability as well as difficulty support this intuition, although the correlations are not as strong as between effort and performance.

The following sections analyse the choices of 157 students who played the Memory game twice with a variable payment scheme, resulting in 314 observations. We motivate this step by comparing the performance of the participants in the two variable payment rounds with the use of the non-parametric Spearman rank correlation coefficient. This test shows that the participants do not change behaviour meaningfully between the rounds (ρ = 0.67, statistically significant at the 1%

[10] For each median regression we performed a corresponding OLS-regression. The calculation of the residuals and a subsequent Shapiro-Wilk test for normality confirms that the error terms are not normally distributed.

level). As the performance levels in these two rounds are not perfectly correlated though, we control for the round with a dummy variable in the following regressions. In one out of the three sessions where the penalties, base wage and bonuses contract (T4) has been employed, we used a different exchange rate. We find that in this treatment there is no effect of the exchange rate on the performance of participants between the rounds of the bonus-penalty contract.[11] Contracts are not economically equivalent because of the flat, performance-independent region in the payment schemes. This is why we employed different exchange rates to ensure similar monetary payments for participants of different treatments. Although we suppose that participants evaluate and decide on their effort provision in terms of points (rather than exchanging the points to money in their mind), we control for this possibility by including the exchange rate in the regression analysis.

Table 2: Descriptive statistics and Spearman's ρ

N=314	Median	SD	Min	Max	1	2	3	4
1 Performance	16	15.60	0	40	1.00			
2 Effort	122.50	90.07	0	341	0.93***	1.00		
3 Ability	22.03	8.34	3.06	42.47	0.19***	0.10*	1.00	
4 Difficulty	7	2.15	1	10	-0.23***	-0.14**	-0.25***	1.00

Notes: Significance levels: *p<0.10, **p<0.05, ***p<0.01

4.2 The effect of payment schemes on effort and performance

In this section we test Hypothesis 1 on the effect of the framed payment schemes on performance. As there is no *direct* causal effect between these two variables, we firstly analyse the treatment effect on effort and subsequently the effect of effort on performance. The payment schemes though are based on performance and so are the cutoff points in the contract frames. Therefore, we assess differences in effort on ranges based on performance which we specify below. Similarly, as we are interested in the effect of the payment schemes on performance (via effort), we look at treatment differences in performance as well (cautiously). The question that is to be answered in this section is how much of the effect of the payment schemes on effort translates to performance and whether there are other variables influencing performance ultimately. We point out differences and similarities when performing non-parametric statistics on both variables. This comparison gives us a hint on which other factors have an impact on performance. To be able to quantify the differences between the effects of the payment schemes, we perform several median regressions.

We suppose that the framed variable payment schemes have a larger effect on effort around the reference point (the performance-independent part), the further away we get from the reference point the less powerful the tool of strategic framing (due to diminishing sensitivity). This means that for small ranges above and below the cutoff points, namely 5 and 10 pairs found, the effect of the payment scheme frame should be the greatest. We therefore choose the ranges 0-5, 0-10 and 0-15 pairs for which we execute the non-parametric tests and regressions. On the contrary, for pairs above 15 up until the maximum amount of 40 pairs, the frames should not have an effect on effort. We expect some of the payment schemes to have equal effects on effort on several of these intervals as the payment in terms of points is the same (we control for the exchange rate in the regressions): the penalties and base wage treatment (T3) should be equal to the contract with penalties, base wage and bonuses (T4) on the range of 0-5 pairs. Similarly, the contract with base wage and bonuses (T2), the penalties and base wage contract (T3) and the contract with penalties, base wage and bonuses (T4) pay a performance-independent amount of 100 points on the range of 6-10 pairs. On the performance range 15-40 pairs, the contract with base wage and bonuses (T2) and the contract with penalties, base wage and bonuses (T4) pay the same amount of points.

Based on these predictions, we first statistically test for the difference in effort between the treatments and afterwards we test for the difference in performance. We test for differences with

[11] Mann-Whitney test, $Prob > |z| = 0.33$.

the Kruskal-Wallis test in the entire range of 0-40 pairs found, and subsequently in the specified intervals of 0-15, 0-10, 0-5 and 6-10 pairs found.

4.2.1 Effort

On the range 0-40 pairs the Kruskal-Wallis test indicates that effort differs significantly across treatments (adjusted χ^2=76.20 with 3*df*, p<0.01). From the descriptive statistics though, we suspect that this difference mainly stems from the penalties and base wage contract treatment (T3). Multiple comparisons with the Kruskal-Wallis test between the treatments confirm this: we reject the null hypothesis for equal distributions of effort for comparisons including the penalties and base wage contract (T3), we fail to reject it for all other comparisons across treatments. The large effect sizes in the first column of Table 3 indicate that the penalties and base wage contract performs significantly worse than all other contracts on the complete performance range.[12] The median regression (see Table 4 for all median regressions) however suggests that the penalties and base wage contract (T3) does not have a significantly different effect on effort on the whole range of performance compared to the base wage and bonuses only contract (T1) when the control variables are included.

We perform several Kruskal-Wallis tests to investigate the differences between treatments on specified intervals for the reasons described earlier: reducing the interval to 0-15 pairs found we observe that the base wage and bonuses only contract (T1) performs significantly better than the penalties and base wage contract (T3) and the contract with penalties, base wage and bonuses (T4). As the contracts are not economically equivalent, we control for the exchange rate in the median regression and find even stronger evidence that the base wage and bonuses only contract (T1) outperforms the other contracts: a participant who is offered a base wage and bonuses only (T1) will, on average, try 67 times to find a pair, compared to only 46 times when offered a base wage and bonuses (T2) and penalties, base wage and bonuses (T4). In the penalties and base wage contract (T3), the participant would only try 32 times.[13]

Table 3: Cohen's d corrected for uneven groups for the four treatments, differences in effort

Treatments	Pairs 0-40	Pairs 0-15	Pairs 0-10	Pairs 0-5	Pairs 6-10	Pairs 15-40
T1 - T2	0.04	0.19	1.14***	0.58	1.19	0.04
T1 - T3	1.57***	1.36***	1.14***	0.50	1.63***	0.04*
T1 - T4	-0.01	0.52*	1.01***	0.64	0.27	-0.10
T2 - T3	1.37***	-0.67	-0.54	-0.35	0.24	0.71*
T2 - T4	-0.05	0.21	-0.15	0.01	-1.50	-0.13
T3 - T4	-1.41***	-0.48	0.34	0.43	-1.50	-0.83*

Notes: Significance based on the Kruskal-Wallis test for differences in distributions
Significance levels: *p<0.10, **p<0.05, ***p<0.01

Reducing the interval further to 0-10 pairs found, we find that the base wage and bonuses only contract (T1) differs significantly from all the other treatment contracts (Table 3). The large effect sizes show that the base wage and bonuses contract outperforms all other contracts significantly. Again controlling for the exchange rate in the median regression strengthens this effect even more: while the treatment with a base wage and bonuses (T2) ranks last with 73 trials, the treatment contract with base wage and bonuses only (T1) reaches the top rank with 114 trials on this range of performance. Surprisingly, the trials in penalties and base wage contract (T3) exceed the ones of the penalties, base wage and bonuses contract (T4). They score 86 versus 82 trials, respectively.

Decreasing the size of the interval to 0-5 pairs found, we cannot find any significant difference between the treatments except for a difference (significant on the 10% level only) between the base

[12] We calculate Cohen's d from pairwise Mann-Whitney tests.

[13] Trials are the measure for effort. We take the median values for age, friends, risk aversion, loss aversion, pleasure for this illustration. Similarly, we take the exchange rate of each treatment, the second round, a male, Dutch, economics student for the calculations. For treatment 4 we take the mean exchange rate of 128.95. When taking the exchange rate of 100, treatment 4 would outperform treatment 2 with 65 trials, when taking the exchange rate of 150, it ranks last with 31 trials.

Table 4: Median regression results with effort as dependent variable

Effort	Pairs 0-40	Pairs 0-15	Pairs 0-10	Pairs 0-5	Pairs 6-10	Pairs 15-40
Treatment						
2	6.80	-55.57***	-86.42***	-40.72*	-47.75	24.43
	(50.01)	(21.10)	(23.56)	(20.33)	(29.35)	(28.08)
3	-115.36	-104.65***	-118.34***	-36.19	-80.73	24.72
	(83.24)	(35.02)	(37.37)	(35.12)	(51.48)	(54.98)
4	32.28	-70.91**	-95.90***	-35.14	-27.83	29.15
	(53.80)	(30.14)	(25.64)	(28.97)	(40.91)	(35.94)
Exchange rate	0.13	-0.69**	-0.90***	-0.16	-0.55	0.24
	(0.81)	(0.34)	(0.34)	(0.31)	(0.50)	(0.46)
Age	-0.81	1.10	1.00	3.49***	-0.39	-2.10
	(1.88)	(0.83)	(0.73)	(0.93)	(0.78)	(2.20)
Male	-0.74	-5.29	-1.76	-6.69	4.50	21.35
	(13.98)	(6.70)	(5.35)	(8.45)	(6.91)	(13.22)
Nationality NL	13.79	-2.39	-1.95	-0.71	0.49	-8.59
	(12.75)	(6.88)	(5.49)	(9.48)	(6.57)	(15.15)
Study						
Law	-73.28	-20.81	-2.34	-12.19	-33.53*	34.88
	(70.93)	(20.10)	(11.65)	(14.05)	(17.81)	(30.78)
Psychology	-18.69	-29.21**	-23.74**	-19.27*	-6.19	15.93
	(29.39)	(14.36)	(11.17)	(11.39)	(19.90)	(35.63)
Medicin	11.21	-23.10*	-19.21*	-5.76	0.07	13.83
	(18.42)	(12.99)	(11.30)	(23.69)	(12.06)	(25.19)
Sociology	-12.41	-7.53	-3.26	29.96**	-13.60	32.96
	(21.37)	(8.31)	(9.30)	(13.45)	(10.18)	(32.09)
Exact Sciences	10.49	-10.61	-10.67	-27.34*	-6.26	14.56
	(16.01)	(13.26)	(15.33)	(14.09)	(10.75)	(20.80)
Humanities	14.52	7.57	6.74	11.64	11.17	-42.30
	(21.80)	(7.52)	(8.16)	(13.20)	(11.25)	(29.85)
Other	3.69	-5.22	-6.69	-17.18*	2.98	22.94
	(14.02)	(11.18)	(9.78)	(10.03)	(7.43)	(19.81)
Friends	2.41	-1.83*	-0.25	-3.69	-0.71	2.33
	(2.72)	(0.98)	(0.80)	(4.76)	(1.24)	(3.86)
Round	-11.26	-12.27**	-5.98	-4.69	2.27	0.38
	(9.94)	(5.17)	(5.97)	(5.60)	(5.72)	(12.17)
Risk aversion	-1.83	1.79	-0.12	-0.01	0.16	-2.50
	(3.03)	(1.38)	(1.60)	(1.41)	(1.67)	(3.23)
Loss aversion	0.81	-5.17	-3.45	-6.59	0.74	14.67**
	(6.91)	(3.45)	(3.09)	(4.37)	(3.54)	(5.94)
Pleasure	3.74	3.55***	2.89**	-2.54*	1.44	-3.01
	(2.72)	(1.28)	(1.36)	(1.36)	(1.47)	(2.89)
Constant	143.06	188.65**	277.99***	20.07	172.04*	163.55
	(187.70)	(80.99)	(62.94)	(77.54)	(100.46)	(122.84)
N	314	152	124	60	64	170
R^2	0.2707	0.2906	0.3432	0.4013	0.3494	0.0944
Robust	Yes	Yes	Yes	No	No	No

Notes: Significance levels: *p<0.10, **p<0.05, ***p<0.01, (robust) standard errors in brackets

wage and bonus treatment (T2) and the penalties and base wage treatment (T3). Including the control variables in the median regression shows that the base wage and bonuses treatment (T2) performs worse than the base wage and bonuses only treatment (T1) with a slight significance only. We argued in the introduction of this section that the penalties, and base wage contract (T3) and the penalties, base wage and bonuses contract (T4) should have the same effect on effort on the performance range of 0-5 pairs. Comparing these two treatments with the Mann-Whitney test supports this (p=0.17).

The results of the Kruskal-Wallis test for the performance range 6-10 pairs, suggest that the treatments with a flat region (T2, T3 and T4) reach significantly similar effort levels, as expected. This is furthermore supported by the median regression on this interval of performance.

When looking at the complete performance range (Table 3), only the penalties and base wage contract (T3) is distinct in terms of effort from the base wage and bonuses only contract (T1). These differences in the effects on effort show in the performance range of 0 to 15 pairs (see median regressions in Table 4) rather than in the upper performance region of 15 to 40 pairs. This indicates that the frame used is most effective around the induced reference point. The median regression points out that participants with higher loss aversion give significantly more effort in the 15-40 pairs range, which we will discuss in Section 5. Against our expectation that only the base wage and bonuses contract (T2) and the penalties, base wage and bonuses contract (T4) should be equal on the range from 15 to 40 pairs, only the penalties and base wgae treatment (T3) has smaller effort levels compared to the other treatments (only on a 10% significance level). This difference does not translate into significant differences in the median regression. From this analysis Hypothesis 1 is supported only on the performance ranges 0-10 and 0-15 pairs.

4.2.2 Performance

Now that we have ascertained the effect of the framed payment schemes on effort in the two performance ranges 0-10 pairs and 0-15 pairs, we focus on how much of this effect translates to performance.[14] For this purpose, next to visual inspection, we again employ a median regression with performance as dependent variable, effort, ability and difficulty as independent variables. We discuss the effect of the latter two variables in the following section. In the median regressions for each performance range (Table 5), we include the main effects and we investigate whether an interaction effect between effort and ability is significantly present. If the interaction term is significant, we use the extended model in the following paragraphs, if not, we retain to the regression with main effects only.[15] We follow the advice of Brambor, Clark, and Golder (2006) to keep all constitutive terms in the regression analysis, even if these turn insignificant when including the interaction term.

For the purpose of testing for differences in the performance levels of the participants in the four treatments, we firstly look at the cumulative percentages of the participants that stopped after each pair. We visualise these for the first variable payment round in Figure 5a and closer for the pairs 0 to 15 in Figure 5b. Three main conclusions can be drawn: firstly, the majority of participants in the penalties and base wage treatment (T3) stop after 5 pairs. Secondly, fewer participants in the base wage and bonuses only treatment (T1) stop up until 10 pairs. Thirdly, from 10 pairs onwards, the base wage and bonuses only contract (T1), the base wage and bonuses contract (T2) and the penalties, base wage and bonuses contract (T4) perform similarly in terms of cumulative stopping percentages.[16]

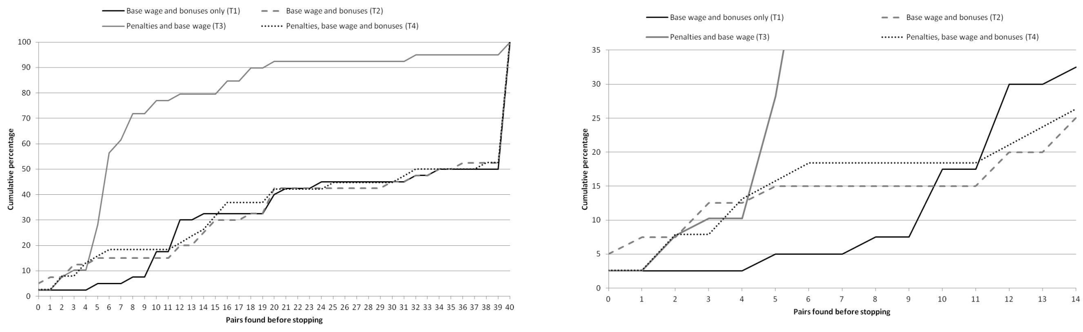
To evaluate the differences in performance numerically, we take the median effort levels of the treatments for the performance ranges 0-10 and 0-15 pairs.[17] We use these number of trials for the calculation of the median performance, with employing median values for ability (21.53 and 21.30 points) and difficulty (7) on these two performance ranges. For the performance range 0-15 pairs, we employ the main effects model: the treatment with a base wage and bonuses only (T1) achieves a median performance of 8 pairs compared to both 5.4 pairs in the treatment with a base wage and bonuses (T2) and in the treatment with penalties, base wage and bonuses (T4). The penalties and base wage treatment (T3) performs worst with a median of 4.7 pairs.

[14] Performing the Kruskal-Wallis test on performance across treatments yields similar results to the ones presented in the previous section with effort (adjusted χ^2=68.30 with $3df$, p<0.01, for details see Table 12 in the Appendix).

[15] On the performance range 0-5 pairs the effects of ability and difficulty are insignificant and therefore we do not consider an interaction effect. The model then reduces to one with effort as only independent variable. See Table 13 in the Appendix for the other specifications of the model.

[16] As discussed in Section 4.1, these effects differ only minimally between the two rounds (see Figure 9a and 9b in the Appendix).

[17] T1: 60 and 70 trials, T2: 21 and 45 trials, T3: 38 and 38 trials, T4: 27 and 45 trials, respectively.



(a) Pairs 0-40 (b) Pairs 0-15

Figure 5: Cumulative stopping percentages for the first variable payment round

For the performance range 0-10 pairs, we employ the main effects model and conclude that the base wage and bonuses only contract (T1) still outperforms all other contracts with 7 pairs found. Interestingly, the penalties and base wage contract (T3) comes in second with a median performance of 5 pairs, followed by the penalties, base wage and bonuses contract (T4) with 4 pairs and finally the base wage and bonuses contract (T2) with 3 pairs. Importantly, the penalties and base wage contract (T3) has the same median effort level for both the performance range 0-10 pairs as for the performance range 0-15 pairs. The same observation can be made when looking at performance which in the median is 5 pairs for both performance ranges. This suggests that the participants do not increase effort (and stop at 5 pairs, ascertained visually in Figure 5a) and therefore lag behind in performance on the range 0-15 pairs compared to participants in the other treatments. In turn this means that higher performance ranges are achieved only with payment schemes that do not only include losses.

Although it is suggested in the literature that penalties can work well to increase effort, this analysis shows that using penalties only is not optimal to achieve higher performance levels. Surprisingly, the base wage and bonuses contract (T2) performs worst on the performance range 0-10 pairs in both calculated median effort and performance, although it pays the same amount of points as the penalties and base wage contract (T3) and the penalties, base wage and bonuses contract (T4) on the range 6 to 10 pairs. It could therefore be a reasonable explanation that the higher performance is an effect of the penalties on the range 0-5 pairs of the latter two contracts. This effect though seems to "wear off" on the performance range 10-15 pairs already, as the penalties and base wage contract (T3) lags behind in effort as well as performance compared to the base wage and bonuses contract (T2) and the penalties, base wage and bonuses contract (T4). Again the latter two contracts achieve the same amount of (calculated) median effort and median performance on the range 0-15 pairs. These results suggest that penalties work well if low performance levels are required. Combining penalties with bonuses (as in T4) achieves superior results for low performance levels than employing only bonuses. However, for higher performance levels this effect disappears. Overall, our results suggest that framing in terms of bonuses only on top of a base wage (as in T1) achieves superior effort and performance levels than any other payment scheme. The non-parametric tests and regression results support the theoretical findings of Hilken et al. (2013) and Hypothesis 1 that strategically framing payments as gains with a low base wage is optimal. The further away we go from the framed reference point, the less the effect of any framed incentive scheme (diminishing sensitivity). Other factors that we will discuss in Section 5 become more important. Similarly, using penalties only is not advised when higher performance levels are required, in this case above 15 pairs. As suggested in the regression in Table 5, ability of the participant and difficulty of the task play an important role when examining the effect of effort on performance. We will discuss these issues in the next sections.

Table 5: Median regression results for performance as dependent variable

Performance	Pairs 0-40	Pairs 0-15	Pairs 0-10	Pairs 0-5	Pairs 15-40
Model	Interaction	Main effect	Main effect	Reduced	Main effect
Effort	0.1397***	0.1033***	0.1036***	0.0926***	0.1250***
	(0.111)	(0.0094)	(0.0116)	(0.0134)	(0.0113)
Ability	-0.0001	0.0725***	0.0693***	-	0.2008***
	(0.0335)	(0.0216)	(0.0253)		(0.0722)
Difficulty	-0.2334**	-0.3858***	-0.2698***	-	-0.7531**
	(0.1176)	(0.1048)	(0.0931)		(0.3145)
Interaction Effort*Ability	0.0009**	-	-	-	-
	(0.0005)				
Constant	0.9854	2.0744**	1.1598	0	8.4459**
	(1.2048)	(0.9663)	(0.8972)	(0.2676)	(3.7662)
N	314	152	124	60	170
R^2	0.8758	0.6716	0.5826	0.5317	0.5717

Notes: Significance levels: *p<0.10, **p<0.05, ***p<0.01, robust standard errors in brackets

4.3 Ability

In the beginning of the experiment we tested the participants as to their ability to remember visuo-spatial information in ten stages of reproducing a pattern. We combined the ten resulting variables into one that is our approximation for individual ability of the participants in the main task. In this section we analyse the effect of this ability variable on performance for all the performance ranges. For first statistical tests, we categorised the participants into three groups with respect to their ability. The lowest 25% are in the first category, the highest 25% in the last category and the rest in the middle category. Before studying the interaction between payment schemes and participants' ability it is important to ensure that there are no significant differences in the treatments as to the ability levels. For a first overview, we summarise the number of participants over the treatments for the low- and high-ability categories in Table 6. At first sight, there are no substantial differences as the percentage of low-ability agents in the treatments ranges between 22.5 and 29 and of high-ability agents between 20 and 31. We perform a Kruskal-Wallis test which supports that ability does not vary significantly across treatments[18], meaning we have an equal distribution of ability of participants across treatments. Furthermore it is important to verify whether the performance of participants with low ability is different from the one of high ability participants. We do so with the Mann-Whitney test for the two categories at each end of the distribution (with both 40 observations for each variable payment round, therefore 80 for the two ability categories). We find that low and high ability participants differ significantly at the 1% level as to their performance in the main task ($z = -3.10$, $p = 0.0019$).

Table 6: High and low ability per treatment

	Number of participants		Total number of participants	%	
Payment scheme	High ability	Low ability		High ability	Low ability
T1	9	9	40	22.5	22.5
T2	8	9	40	20	22.5
T3	12	11	39	30.77	28.21
T4	11	11	38	28.95	28.95

Notes: Top and bottom 25% of all; Low ability 3.06 to 16.48 points and high ability 29.57 to 42.47 points

The median regression run to evaluate the effect of effort on performance also includes the ability variable (Table 5): this tells us more about the magnitude of the relationship between

[18]Based on 157 observations, χ^2 corrected for ties=0.44 with 3 d.f. and p=0.93

ability, effort and performance and allows us to evaluate Hypotheses 2a and 2b. In Table 7 we calculate performance with median effort and effort with median performance for different levels of ability based on the median regression on the performance range 0 to 40 pairs.

To see the change in performance and effort with increasing ability, we took the minimum and maximum points in the classifications of both low and high ability and calculated the estimated values of performance and effort according to the median regression. We suspected that higher-ability agents reach higher performance levels compared to low-ability agents, given the level of effort. The first row in Table 7 shows that with the median effort of 122.5 trials, a low-ability agent would accomplish finding between 17 and 18 pairs. The high-ability agent in contrast to this would achieve a performance between 20 and 21 pairs. The difference only based on different ability levels lies between 2 and 4 pairs, which corresponds to a 5 to 10% difference based on the maximum performance level of 40 pairs. This denotes a significantly higher productivity for high-ability agents and therefore supports Hypothesis 2a.

Table 7: Performance and effort calculations for minimum and maximum points in high and low ability classification

	Ability in points				
	Min 3.06	25th perc. 16.48	Median 22.03	75th perc. 29.57	Max 42.47
Hypothesis 2a: Performance in pairs with median effort=122.5 trials	16.80	18.28	18.89	19.72	21.14
Hypothesis 2b: Effort in trials with median performance=16 pairs	116.87	107.74	104.37	100.12	93.59

The next question to answer is whether high-ability agents need less trials for the same performance outcome compared to low-ability agents. The results in the second row of Table 7 are obtained using the same method as in the first row. For a high-ability agent to find 16 pairs (the median performance), it takes between 94 and 100 trials in contrast to the low-ability agent who needs to try between 108 and 117 times. Low-ability agents need a significantly greater amount of effort for the same performance level, concluding that Hypothesis 2b is supported as well.

In the main effects specification of the median regression, ability (difficulty) has a significantly positive (negative) effect in all performance ranges, except in the range of 0 to 5 pairs. We subsequently exclude the variables from the regression analysis on this performance range and observe that effort is the only explanatory variable. When the performance range is increased to include all agents up to 10 pairs, both ability and difficulty become significant. This means that with increasing performance levels, ability and difficulty have, next to effort, an additional effect on performance. While the effect of effort roughly stays the same when enlarging the range to 0-15 pairs, the effects of both ability and difficulty increase. Further increasing this range to 40 pairs, we observe that the larger the ability of the agent, the larger the effect of effort on performance (the interaction effect of ability and effort is significantly positive). However, this interaction effect is only present on the whole performance range. diminishes when looking at the complete performance range, which can be explained by the fact that the interaction is not significant on the upper performance range of 15 to 40 pairs. For agents that are reacting to the frames[19], those that stop between 0 and 15 pairs, ability has a multiplicative effect with effort on performance. This effect is not present any longer for agents that stop above 15 pairs. We depicted the overall effect of effort on performance for different ability levels (minimum, median and maximum) in Figure 6.

To evaluate whether the different payment schemes have different effects on high- and low-ability agents, we make use of both the median regressions performed on the range 0 to 15 pairs.[20]

[19]See Table 4, the effect of the contract frames is significant on the ranges 0 to 10 pairs and 0 to 15 pairs.

[20]On the complete performance range we see that the treatments do not have an effect, therefore the choice of performance range. We could similarly use the range of 0 to 10 pairs. The reason to use a prediction of values is that it makes the performance of high- and low-ability agents comparable across the treatments as it includes all other effects as well.

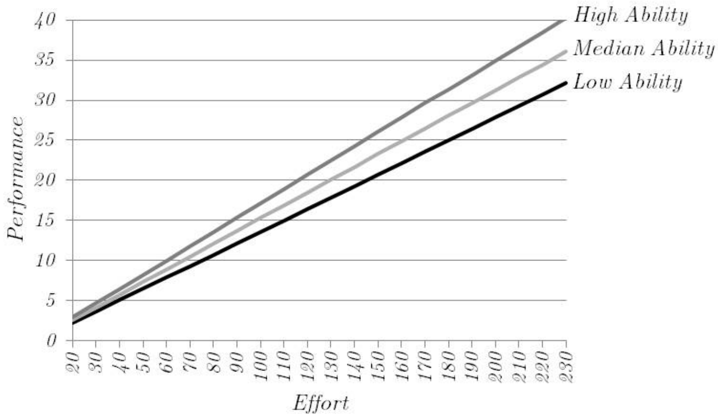


Figure 6: Effect of effort on performance for different ability levels

We firstly predict the effort provided by all the agents as if they stopped on the performance range 0-15 pairs (based on the median regression in Table 4). Secondly, we take these individual values for effort and predict their performance with the help of the median regression on the range 0-15 pairs in Table 5. We subsequently perform several Mann-Whitney tests to evaluate differences in performance across treatments for high- and low-ability agents. We equally compare the performance of high-and low-ability agents in the treatments. In the base wage and bonuses only contract (T1), the base wage and bonuses contract (T2) and the penalties and base wage contract (T3) we see significant differences between high- and low-ability agents (see Table 14 in the appendix for all tests and data).[21] We calculate Cohen's d to be able to make the effects comparable across treatments: the largest difference in performance shows in the base wage and bonuses only contract (T1) with an effect size of -1.88, followed by the base wage and bonuses contract (T2) with an effect size of -1.00. The differences in performance between high- and low-ability agents is smallest in the penalties and base wage contract (T3, effect size: -0.64). The combination of bonuses and penalties eliminates the effect of ability: in the penalties, base wage and bonuses contract (T4) there are no significant differences between the two ability groups. Furthermore, we find that low-ability agents perform similarly in three treatments: the base wage and bonuses contract (T2), the penalties and base wage contract (T3) and the penalties, base wage and bonuses contract (T4). In the base wage and bonuses only contract (T1) though, low-ability agents perform significantly better than in all other treatments. For high-ability agents this is true as well. The main difference lies in the fact that high-ability agents perform significantly worse in the penalties and base wage contract (T3) compared to all other contracts. The base wage and bonuses contract (T2) and the penalties, base wage and bonuses contract (T4) achieve similar performance levels of high-ability agents. Indeed, the contracts that include bonuses (T1 and T2) induce higher performance levels for high-ability agents compared to low-ability agents, which supports Hypothesis 3a. Framing penalties only in a contract (T3) induces lower performance levels of low-ability agents compared to high-ability agents. This again supports Hypothesis 3b, but with a side-note: using penalties (T3) achieves similar performance levels compared to both bonuses (T2) and the combination of penalties and bonuses (T4). The differences between high- and low-ability agents performance indeed vanishes when penalties and bonuses are combined (consistent with Hypothesis 3c). Interestingly, high-ability agents perform worse when only penalties (T3) are used, but not when they are combined with bonuses (T4). Similarly, when penalties are used for low performance ranges, this does not have an effect on the performance (comparison of T2 and T4). We can conclude that indeed high-ability agents react more (positively) to bonus frames, low-ability agents react more (negatively) to penalty schemes in terms of performance and that

[21] T1: $p<0.01$, T2: $p<0.05$, T3: $p<0.05$, T4: $p=0.11$

the two effects combined cancel each other out.

4.4 Performance and cost

Up to this point we have disregarded the principal, and therefore the cost side of the contractual relationship. Eventually it is the principal who has to pay the agents for the output they created. Equivalently, it is the principal that determines the total output that is required. In this section we consider how many heterogenous agents the principal has to employ under each of the payment schemes, the respective performance per agent, the costs associated with the total production, as well as per unit costs on several performance ranges. As the experimenter is the principal in this case, we look at how much money we have spent per treatment for some pre-specified capacity, and how much each single unit costs. From this it is then possible to determine which payment scheme is minimising the costs of the principal.

We firstly look at the average number of units each agent produces under the different payment schemes on increasing performance ranges (see Table 8). To this end, we take the amount of units produced of all the agents stopping on that performance range and divide this by the number of agents. It shows that agents in the base wage and bonuses only treatment (T1) have the highest average number of products on all but the lowest performance range. On the lowest range though we see that the penalties and base wage treatment (T3) performs better. This implies in turn that the principal needs fewer agents for the respective amount of units for the agents employed under these payment schemes. Suppose now, that each of the agents in each of the performance ranges produces the average number of units. As we have used different exchange rates in each of the treatments, the costs per agent are not equivalent to the points they have received. We calculate the number of points each agent would receive for the production of the average number of units and subsequently we translate this to Euro amounts. The total cost of production is the number of agents times the per agent payment in Euros, the unit cost is derived by dividing the total costs with the number of units.

In Table 8 we observe that for the total range of 0-40 pairs found the penalties and base wage treatment (T3) has similar total costs as the base wage and bonuses only treatment (T1) for the production of the same amount of units. The majority of agents in the penalties and base wage treatment (T1) though stop producing after an average of 5.6 units (85.9% stop on the range 0-10 pairs, see Table 19 in the Appendix). For this reason the principal needs to employ more agents to produce the same amount. A reasonable criticism would be to claim that setting the cutoff point in the penalties and base wage contract at 5 pairs has a significant impact on the high stopping percentage in the first quarter of maximum production. Increasing the base wage and the cutoff point would then be a solution which comes at a price though: increased costs. The penalties and base wage contract (T3) would have higher per agent costs and subsequently also total and unit costs. This makes this contract less efficient than the base wage and bonuses only contract (T1). The least costly contract is the one where a base wage is combined with bonuses for higher performance ranges only (T2), both in total as well as per unit for the higher performance ranges. For the lower performance ranges, the penalties, base wage and bonuses contract (T4) has the lowest total and unit costs.

This analysis suggests that using only penalties in the frame of a payment scheme does not reduce the costs of production because of the lower average per agent production. Both the penalties and base wage contract (T3) and the penalties, base wage and bonuses contract (T4) are more costly than using bonuses on top of a base wage for higher performance levels (T2) to the principal. Although the penalties, base wage and bonuses contract (T4) achieves similar average production levels per agent on the whole performance range compared to the contracts including only bonuses (T1 and T2), this comes at a higher cost for the principal. These results support Hypothesis 4, but differently than expected: employing penalties only is not cost-minimising, not even for lower performance levels. Employing bonuses for higher performance levels is optimal with respect to costs, employing a frame with penalties, a base wage and bonuses is cost-effective for lower performance levels.

Table 8: Average production of agents stopping on the specified ranges in the different treatments

Payment scheme	Average production	Number of agents	Costs in points per agent	Costs in Euros per agent	Total cost of production	Per unit costs
0-5 pairs, Maximum capacity: 250 units						
T1	2.80	90	156	0.78	70.20	0.28
T2	1.60	157	100	0.67	105.19	0.42
T3	**3.44**	**73**	69	0.69	50.37	0.20
T4	2.07	121	**42**	**0.34**	**41.14**	**0.16**
0-10 pairs, Maximum capacity: 500 units						
T1	**7.53**	**67**	251	1.26	84.42	0.17
T2	2.89	174	100	0.67	116.58	0.23
T3	5.60	90	100	1.00	90	0.18
T4	2.89	174	**58**	**0.46**	**80.04**	**0.16**
0-20 pairs, Maximum capacity: 1000 units						
T1	**10.89**	**92**	318	1.59	146.28	0.15
T2	9.27	108	**100**	**0.67**	**72.36**	**0.07**
T3	6.64	151	**100**	1.00	151.00	0.15
T4	8.62	117	**100**	0.80	93.60	0.09
0-30 pairs, Maximum capacity: 1500 units						
T1	**12.73**	**118**	355	1.78	210.04	0.14
T2	11.56	130	132	0.88	**114.40**	**0.08**
T3	6.64	226	**100**	1.00	226	0.15
T4	9.95	151	**100**	**0.80**	120.80	0.08
0-40 pairs, Maximum capacity: 2000 units						
T1	**25.85**	**78**	617	3.09	241.02	0.12
T2	24.49	82	390	2.60	**213.20**	**0.11**
T3	8.24	243	**100**	**1.00**	243.00	0.12
T4	24.79	81	396	3.17	256.61	0.13

Notes: Highest scoring value on each range in bold. Based on the two variable payment rounds combined. The calculated number of agents is rounded to whole numbers. We take the averages for the two different exchange rates in treatment 4 as the exchange rate does not have an effect on performance as shown earlier.

5 Discussion

In this section we discuss several issues that are not included in the main research questions of the paper, but that are relevant in the context of reference-dependent preferences and framing of payment schemes.

Difficulty: In the experiment we asked the participants to self-report their perceived level of difficulty This variable has a significantly negative effect on all performance ranges, except for the lowest (0-5 pairs). We conjecture that the principal could consider to frame the difficulty of the main task, e.g. making participants *perceive* the main task easier. If this framing works, the principal would gain between 1.4% and 3.9% in median performance with decreasing perceived difficulty by one point. This conjecture has to be confirmed in future research.

Risk aversion: In the experiment we elicited the participants' level of risk aversion. If the main task includes a lot of chance, e.g. finding pairs by accidentally turning two identical cards, participants' risk aversion may affect their performance on the task. As we are reshuffling the deck after the participant found two pairs, the number of trials before finding one pair in the odd rounds gives us an indication of chance that is involved in the game. When the participant needed only one trial to find a pair the first time he sees the deck, this can be attributed to chance. Only 2% of the total number of pairs found can be traced back to chance, therefore the insignificance of risk aversion in the median regression is not surprising.[22]

[22]If we include the second round as well when the participants saw some of the cards already and only needing one trial might be explained by ability, this percentage only increases to 5%. It cannot be excluded that risk does not play a role in the game at all, but these percentages are a strong support for sheer luck to be minimal in the main task.

Loss aversion: Brooks, Stremitzer, and Tontrup (2012) examine why loss frames might have a positive effect on the effort provision of the agent and recommend loss aversion as explanatory variable. In our experiment we interestingly find that loss aversion only increases effort for the participants who stop on higher performance levels and not on the ranges where we find treatment effects. One explanation might be provided by Brooks et al. (2012, p.81) who identify two channels through which loss framing works: "[...] the loss frame influences subjects' effort choices [...] by an endowment effect and by setting a default expectation that cognitively induces subjects to invest more effort.". They hereby suggest that agents facing a loss frame interpret the threshold value as the performance level which the principal expects them to deliver, rather than being motivated by individual loss aversion.

Reference point / Framing of the base wage: The most sensitive point of the experimental design is the question whether setting a base wage in fact induces a reference point for the participants. Investigating what the reference point is, is extremely difficult and does not only pose challenges for our experimental design. As discussed by De Meza and Webb (2007), there is no agreement in the literature about the nature of the reference point. In this experiment we (assumed to) induce a reference point by paying participants a base wage for two rounds and keeping this base wage in the variable payment scheme rounds. An alternative to this specification is to assume that the reference point is induced by expectations, as described by Kőszegi and Rabin (2006). To get further insight into the formation of the reference point in our experiment, we asked the participants between the two variable payment rounds how many points they expect to receive in the following round. As an approximation for the past experience of participants we use the points earned in the preceding round - the first variable payment round. As we are not looking at differences across treatments in this section, we use the data for the two variable payment rounds separately. If expectations determine the reference point, participants should match their performance in the second round with their expectations (as described in the concept of personal equilibrium, see Kőszegi & Rabin, 2006). If the participant stays below his reference point, he feels a loss, which can be avoided by exerting more effort and subsequently finding more pairs. 51% of the participants behave in this fashion. We exclude the participants that find 40 pairs, as these participants might see the maximum as reference point and any performance below this amount as a loss. When taking the treatments that have a flat part in the payment scheme (T2, T3 and T4), 77% of participants stop at the base wage. 58% of the same group of participants match their performance with their expectations. From these observations we conclude that the base wage is a good approximation for the reference point in this experiment, but we cannot exclude that expectations have an influence on the decision-making process of the participants.

In the questionnaire participants were asked to choose between three payoff-equivalent contracts according to their preferences. One contract had the form of a low base wage and bonuses, one contract had the form of a high base wage and penalties and one combined a medium base wage with both penalties and bonuses. 40.76% of the participants indicated that they prefer the bonus contract compared to only 15.29% who would choose to work under an equivalent penalty contract. These preferences are in line with the findings of Luft (1994), who did not evaluate an equivalent bonus/penalty contract in her research: interestingly, the majority of the participants (43.95%) chose the combination of bonuses and penalties as the most preferable contract. These preferences might be important in the design of payment schemes when we take the theory of crowding out of intrinsic motivation into account (see next paragraphs).

Pleasure in task / Intrinsic motivation: Note that in the design of this experiment the task is not boring as in many other experimental designs (Abeler et al., 2009). We believe that requiring the task to be boring does not reflect reality: in general, people do not only dislike their jobs, but they like different features about it. To control for the effect of pleasure experienced from the task, we asked participants in the questionnaire to self-report their experienced pleasure. In the median regression analysis (see Table 4) we find that pleasure has a positive significant effect only on those ranges of performance where the frames of the payment schemes are significantly present as well (i.e. pairs 0 to 10 and 0 to 15). The size of the coefficients (2.89 and 3.55, respectively) suggests that the agent will exert significantly more effort the more pleasure he experiences from the task. With the median pleasure of 6 for both performance ranges this means that the agent tries between 17 and 21 times more. This is an important factor to consider when employing an agent, suggesting that making the task more pleasurable for the agent increases his effort.

Incentive payments, bonuses and/or penalties, might work as positive or negative reinforcers of the agents' behaviour (see Bénabou and Tirole (2003), Frey and Jegen (2001) and Deci, Koestner, and Ryan (1999) for more in depth discussion): extrinsic motivation might crowd out or crowd in intrinsic motivation. As we are using both bonuses and penalties and the combination of the two in the treatments, positive or negative reinforcement could be one effect that is at work when framing. To get an indication for this, we take the mean performance of the agents in the second fixed payment round and the immediately following first variable payment round:[23] In the fixed payment round, where purely self-interested, rational agents should not work at all, performance in all treatments ranges between 16 and 24 pairs. When employing only penalties, performance decreases to only 10 pairs on average. In the other treatments when we employ bonuses (as well), performance increases by 3 to 9 pairs. The treatment in which we combine framing both bonuses and penalties has the greatest effect. Using bonuses only seems to be a less effective positive reinforcer than combining bonuses with penalties. This short description suggests even more that framing payment schemes as penalties can be harmful for the principal. In this case, the principal achieves lower performance and higher costs per unit utilising penalties.

6 Conclusion

Our real effort experiment allows us to deduce several implications regarding the way payment schemes can increase effort and subsequently performance. The results suggest that correctly designing payment schemes for agents enhances the effort an agent is giving in a task and, taking his ability and the difficulty of the task into consideration, increases performance. We now summarise the main findings and hint at possible future alleys for research.

Strategically framing payments to an agent as bonuses only on top of a base wage significantly increases effort and in turn also performance. This way of expressing payments to the agent is superior to any other payment frame: when employing penalties in a payment scheme, effort falls behind the one when only bonuses are framed. On a low range of performance levels, penalties work efficiently in increasing effort, but as soon as the threshold level is reached, agents stop working immediately. When framing penalties in combination with bonuses for higher performance levels, the performance increases, but reaches the same level as when employing an equivalent payment scheme without the penalties for the low performance levels. As suggested in the discussion section, using penalties might also crowd out intrinsic motivation. This is especially important when agents are employed for more than one term. The higher unit costs with bonus schemes could be outweighed by higher motivation in the next term.

In the analysis of our ability variable it turned out that effort and the ability of an agent are complements in achieving performance. This indicates that training employees specifically for their tasks (i.e. increasing their ability) makes the agents' effort more effective. The recommendation is to invest in ongoing training and specialising agents in their tasks. This goes hand in hand with the perceived difficulty of a task by the agent: with specific training, the task at hand becomes easier for the agent and this in turn increases performance. Logically, in this respect we should not forget the effect of experience in decreasing perceived difficulty either. Making a task seem less difficult for the agent might additionally augment to this effect.

Another important factor in recent literature on payment schemes is loss aversion. In our real effort experiment loss aversion does not play a significant role in determining the effort of the agent. Only for agents that go "beyond" the frame, to the upper performance levels, higher

[23] Mean performance rounded to whole numbers, for the second fixed payment round and the first variable payment round, respectively: treatment 1: 24 and 27, treatment 2: 19 and 27, treatment 3: 16 and 10, treatment 4: 17 and 27.

loss aversion increases effort. In these cases, the payment scheme becomes irrelevant, meaning that irrespective of the payment scheme, agents tend to finish the task because of loss aversion. Interestingly, where the frames of the payment schemes have an effect, loss aversion has no effect.

One important aspect for the principal is how to minimise the costs of a contract with an agent. We found that employing a base wage with bonuses minimises the costs for the principal if he wants agents to produce higher levels of output per person. If he requires lower per agent performance, employing a penalties, base wage and bonuses contract is cost-effective. Although penalties are less costly to the agent, the low per agent production requires a larger number of agents to be employed for a target level of output. Employing only bonuses is not cost-minimising.

References

Abeler, J., Falk, A., Götte, L., & Huffman, D. (2009). *Reference Points and Effort Provision.* CESifo Working Paper No.2585. (Category 4: Labour Markets)

Armantier, O., & Boly, A. (2012, April). Framing of Incentives and Effort Provision. *Mimeo*.

Bénabou, R., & Tirole, J. (2003, July). Intrinsic and Extrinsic Motivation. *The Review of Economic Studies*, *70*(3), 489-520.

Berch, D. B., Krikorian, R., & Huha, E. M. (1998, December). The Corsi Block-Tapping Test: Methodological and Theoretical Considerations. *Brain and Cognition*, *38*(3), 317-338.

Bonner, S. E., & Sprinkle, G. B. (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.

Borghans, L., Heckman, J. J., Golsteyn, B. H., & Meijers, H. (2009). Gender differences in risk aversion and ambiguity aversion. *Journal of the European Economic Association*, *7*(2-3), 649-658.

Brambor, T., Clark, W. R., & Golder, M. (2006, Winter). Understanding Interaction Models: Improving Empirical Analyses. *Political Analysis*, *14*(1), 63-82.

Brink, A. G., & Rankin, F. W. (2011). The Effects of Risk Preference and Loss Aversion on Individual Behavior under Bonus, Penalty, and Combined Contract Frames. *American Accounting Association, AAA 2009 Management Accounting Section (MAS) Meeting*.

Brooks, R. R., Stremitzer, A., & Tontrup, S. (2012, January). Framing Contracts: Why Loss Framing Increases Effort. *Journal of Institutional and Theoretical Economics*, *168*(1), 62-82.

Croson, R., & Gneezy, U. (2009). Gender Differences in Preferences. *Journal of Economic Literature*, *47*(2), 448-474.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin*, *125*(6), 627-668.

De Meza, D., & Webb, D. (2007, March). Incentive Design under Loss Aversion. *Journal of the European Economic Association*, *5*(1), 66-92.

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

Fehr, E., Klein, A., & Schmidt, K. M. (2007). Fairness and Contract Design. *Econometrica*, *75*(1), 121-154.

Fehr, E., & Schmidt, K. M. (2007, May). Adding a Stick to the Carrot? The Interaction of Bonuses and Fines. *The American Economic Review*, *97*(2), 177-181.

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

Frey, B. S., & Jegen, R. (2001). Motivation Crowding Theory. *Journal of Economic Surveys*, *15*(5), 589-611.

Fryer, R. G. J., Levitt, S. D., List, J., & Sadoff, S. (2012, July). Enahncing the efficacy of teacher incentives through loss aversion: A field experiment. *NBER Working Paper No. 18237*. (http://www.nber.org/papers/w18237)

Gächter, S., Johnson, E. J., & Herrmann, A. (2007, July). Individual-Level Loss Aversion in Riskless and Risky Choices. In *IZA Discussion Paper Series.* The University of Nottingham.

Hannan, R. L., Hoffman, V. B., & Moser, D. V. (2005). Bonus versus penalty: does contract frame affect employee effort? In (Vol. 2, p. 151-169). Springer US. (In Experimental Business Research)

Herweg, F., Müller, D., & Weinschenk, P. (2008, October). The Optimality of Simple Contracts: Moral Hazard and Loss Aversion. *Bonn Econ Discussion Papers, Bonn Graduate School of Economics (BGSE), University Bonn*, *2008*(17).

Hilken, K., De Jaegher, K., & Jegers, M. (2013, March). *Strategic Framing in Contracts: Contracts under Hidden Action.* Tjalling C. Koopmans Research Insitute, Discussion Paper Series Nr. 13-04. (www.uu.nl/rebo/economie/discussionpapers)

Holmstrom, B. (1979). Moral Hazard and Observability. *The Bell Journal of Economics*, *10*(1), 74-91.

Holt, C. A., & Laury, S. K. (2002, Dec.). Risk Aversion and Incentive Effects. *American Economic Review*, *92*(5), 1644-1655.

Hossain, T., & List, J. A. (2009, October). The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations. *NBER Working Paper No. 1562*.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263-291.

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

Kühberger, A. (1998, July). The Influence of Framing on Risky Decisions: A Meta-Analysis. *Organizational Behavior and Human Decision Processes*, *75*(1), 23-55.

Lazear, E. (2000, December). Performance Pay and Productivity. *American Economic Association*, *90*(5), 1346-1361.

Luft, J. (1994). Bonus and penalty incentives. Contract choice by employees. *Journal of Accounting and Economics*, *18*(2), 181-206.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How Are Visuospatial Working Memory, Executive Functioning, and Spatial Abilities Related? A Latent-Variable Analysis. *Journal of Experimental Psychology: General*, *130*(4), 621-640.

Shearer, B. (2004, April). Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment. *The Review of Economic Studies*, *71*(2), 513-534.

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, *211*(4481), 453-458.

7 Appendix

7.1 Instructions

The instructions for each treatment are given in a separate envelop to the participants. The payments are expressed in the same manner, first naming the reference income and then the gains or losses the participants can influence with their performance.[24]

- Instructions -

Thank you for participating!

These instructions are the same for all participants. Please read them carefully. It contains all the information you need to complete the experiment. If you have any questions, raise your hand, and one of the instructors will answer your question. Please switch off your mobile phone

[24] Due to an operational problem, we have lost data for the first round when subjects received a fixed payment for 100 subjects. We let the subjects play this round twice. For those treatments where the data is available, we tested whether the choices in these two identical rounds is correlated: the Spearman rank correlation between the number of pairs found in the two fixed payment rounds is significant at the 1% level with $\rho = 0.46$.

and do not talk to other participants during the experiment. You are not allowed to write down anything. If you break the rules, you are excluded from the experiment.

This experiment will take you about 60 minutes.

In this experiment you will be asked to carry out several exercises for which you can earn money. For each part you will receive a number of points that are translated into your payment at the end of the experiment at the exchange rate of:

200 points = 1 Euro

Your payment and your answers in the experiment stay completely anonymous. The computer stores all the information you have given for analysis only. Your payment depends on your own decisions and actions. On your table you see two envelopes. The computer will tell you when to use them. Please do not open them now. You will get the time to read the instructions before you start the round.

- Overview of the experiment -

The experiment starts with a test of your short-term memory, consisting of *ten rounds*. The test is followed by *one trial round* of the exercise which is not relevant for your payment, but for you to get used to it. Then, the experiment consists of *four rounds* that you are paid for. After this exercise, *two rounds* will follow with decisions on lotteries. Finally, the experiment ends with a *questionnaire*. The last screen will show you your earnings for each part. We will round your payment up or down to 10 cents for convenience.

Between these rounds, you have to answer two questions. The answers to the questions have no impact on your payments, you are asked to make sure that you understood the points you receive in each round. If you give the right answers, you may start the corresponding round. If you give a wrong answer to (one of) the question(s) you will be asked to read the instructions again and to fill in the right answer(s).

There will also be a question on your expectations. There are no right or wrong answers in this case.

- The short-term memory test -

In this part of the experiment we will test your short-term memory in ten rounds. You receive *10 points* per correct answer in the test. At the end of the experiment the computer will tell you how well you did on this test.

The first screen will show you 9 squares (3 times 3). 3 of these 9 grey squares turn red for 1 second. You have to remember the location of the three red squares. After the red squares have disappeared, you have to click on the grey squares that were red before. The ones that you have clicked on, will then turn back to red. Be careful, you can not change your choice once you have clicked on a square. There is no time-out, so you have time to think about your choice. After you have clicked on three squares, the next round follows. The following round has 12 squares (3 times 4) and you have to find 4 red squares. In the rounds that follow, the number of grey and red squares increases. For an example of what you are going to see, have a look at Figure 1 below.

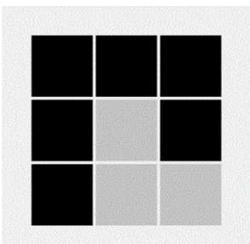


Figure 7: Example of the short-term memory test

- Your task -

In this part of the experiment, we will ask you to play a game: Memory. You will play this scenario for two rounds. The game is the same for each round you are playing. The difference between the rounds is how many points you can earn. You do not play against the computer. The screen will show you 40 cards face down. The deck contains 20 different pictures, two times the same card. The figure below shows you the screen you will see at the beginning of each round. Your task is to find the matching cards by flipping over two cards at a time. You can try to find pairs as long as a total number of 40 pairs have been found. Then the computer will show you an empty screen. Push the *Stop* button to proceed to the next round. If you have found two pairs, the Memory game starts from the beginning. The pairs you will find, add up to the two you have found before. If you don't want to play anymore, you have to push the *Stop* button. Then this round ends. In total there are four rounds. Below the Memory game you will see information on the number of times you tried to find pairs and how many pairs you have found. The number of pairs found in each round determines your points.

- Example -

Note: This is an example. The payments in the paid experiment are not the same!
Suppose that you can earn points in the following way (per round - until you push the *Stop* button):

You receive 100 points. You lose 5 points for each pair below 10. You gain 5 points for each pair above 12.

This means that if you have found *5 pairs*, you receive *75 points* for this first round. If you have found *11 pairs*, you receive *100 points* for the second round. Similarly, if you have found *15 pairs*, you receive *115 points* for the third round. The total of these three rounds would then add up to *290 points*. Your points are exchanged to Euros at the end of the experiment.

- Lotteries -

After you have played the Memory game, there will be two rounds where you need to take decisions.

The first lottery: Your screen will show you 10 rows. In each row, two options are displayed: Option A and B. You need to decide which of the two options you prefer. After the experiment, the computer will randomly pick one of the 10 rows. For that row, the computer then randomly determines your money earnings for the Option (A or B) you chose.

The second lottery: Your screen will show you 6 rows. Each row shows you one lottery. You have to decide whether you accept or reject the lottery. After the experiment, the computer will randomly select one of the rows in the table below. If you have accepted the lottery in that row, the computer will then toss a virtual coin to determine your payoff. If you reject a lottery and the computer chooses that row, your payoff is automatically set to 0.

The outcomes of these two lotteries will be displayed at the end of the experiment.

- Questionnaire -

After you have completed the short-term memory test, the four rounds of Memory and the two lotteries, you are asked to fill in a short questionnaire. Please take your time and fill it in truthfully. The answers you give have no impact on your payments, but they are important for our scientific analysis. In the meantime, your points will be exchanged into your payment. Please stay seated until the payment has taken place.

You will receive all further instructions via the computer.

You can start the experiment by pushing the *Start* button on your screen.

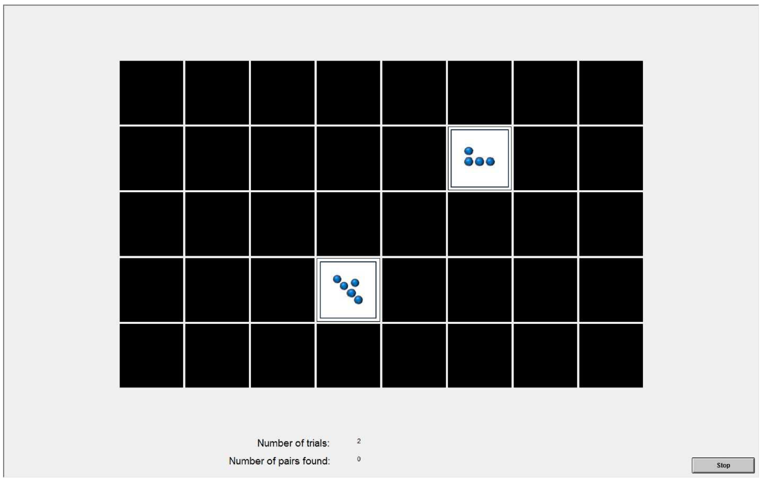


Figure 8: The Memory Game

The following instructions are communicated via cards on the tables and the computer.

- **Card fixed payment**

  Your payment per round: **You receive 100 points.**

- **Card Treatment 1**

  Your payment per round: **You receive 100 points. You gain 20 additional points per pair found.**

  The questions determining whether they have understood the instructions: *"How many points do you receive if you find 10 pairs?"* The correct answer is 300 points. *"How many points do you receive if you find 6 pairs?"* The correct answer is 220 points.

- **Card Treatment 2**

  Your payment per round: **You receive 100 points. You gain 20 points for each pair above 10.**

  The questions determining whether they have understood the instructions: *"How many points do you receive if you find 11 pairs?"* The correct answer is 120 points. *"How many points do you receive if you find 7 pairs?"* The correct answer is 100 points.

- **Card Treatment 3**

  Your payment per round: **You receive 100 points. You lose 20 points for each pair below 5.**

  The questions determining whether they have understood the instructions: *"How many points do you receive if you find 6 pairs?"* The correct answer is 100 points. *"How many points do you receive if you have found 3 pairs?"* The correct answer is 60 points.

- **Card Treatment 4**

  Your payment per round: **You receive 100 points. For each pair below 5, you lose 20 points. For each pair above 10, you gain 20 points.**

  The questions determining whether they have understood the instructions: *"How many points do you receive if you find 4 pairs?"* The correct answer is 80 points. *"How many points do you receive if you find 15 pairs?"* The correct answer is 200 points.

7.2 Questionnaire

- What is your age?
- Are you male or female?
- Where do you come from? (Answer possibilities: The Netherlands, Germany, United Kingdom, Romania, Bulgaria, France, Spain, Other.)
- How many people who also participate in this experiment do you know by first name?
- What do you study? (Answer possibilities: Economics, Law, Psychology, Medicine, Sociology, Exact Sciences (Mathematics, Chemistry, Physics, etc.), Humanities, Other.)
- How pleasant was the Memory game for you on a scale of 1 to 10?
- How difficult did you find the exercise on a scale of 1 to 10?
- Your boss wants you to play the Memory game in this experiment. He gives you the choice between three contracts. Please indicate which one you would choose:

  Option 1: "You are given 100 points. You keep these points if you are able to find exactly 10 pairs. You lose 5 points for every pair below 10, and you win 5 points for every pair above 10."

  Option 2: "You are given 50 points. You win 5 points extra for every pair."

  Option 3: "You are given 150 points. You keep these points if you are able to find exactly 20 pairs. You lose 5 points for every pair below 20."

7.3 Overview of sessions

Table 9: Overview of sessions

Session number	Number of participants	Treatment number
2	16	5
3	17	4
4	15	2
5	17	2
6	16	3
7	15	4
8	20	3
9	17	5
10	12	2 and 4
11	13	2, 3, 4 and 5

Table 10: Category codes and number of participants for nationality and study

Code	Country of origin	N	Study	N
1	The Netherlands	84	Economics	64
2	Germany	6	Law	8
3	United Kingdom	1	Psychology	7
4	Romania	2	Medicin	10
5	Bulgaria	8	Sociology	10
6	France	0	Exact Sciences	14
7	Spain	0	Humanities	13
8	Other	57	Other	32

7.4 Statistics on performance

Table 11: Mean differences of performance for the four treatments

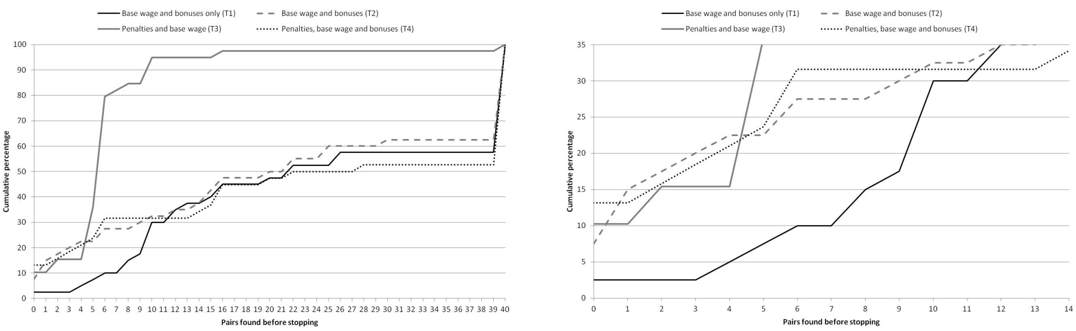
Treatments	Pairs 0-40	Pairs 0-5	Pairs 0-10	Pairs 0-15	Pairs 6-10	Pairs 15-40
T1 - T2	1.36	1.2	4.64***	2.62**	1.46	1.68
T1 - T3	17.61***	-0.64	1.93	3.59***	2.33***	9.28
T1 - T4	0.74	0.73	4.64***	3.4***	3.21*	0.9
T2 - T3	16.25***	-1.84*	-2.71***	0.97	0.87	7.6
T2 - T4	-0.62	-0.47	0	0.78	1.75	-0.78
T3 - T4	-16.87***	1.37	2.71***	-0.19	0.88	-8.38

Notes: Significance based on the Kruskal-Wallis test for differences in distributions
Significance levels: *p<0.10, **p<0.05, ***p<0.01

Table 12: Cohen's d corrected for uneven groups for the four treatments, differences in performance

Treatments	Pairs 0-40	Pairs 0-5	Pairs 0-10	Pairs 0-15	Pairs 6-10	Pairs 15-40
T1 - T2	0.09	0.67	1.49***	0.54**	1.08	0.19
T1 - T3	1.52***	-0.31	0.75	1.25***	1.69***	1.05
T1 - T4	0.05	0.36	1.64***	0.75***	3.02**	0.10
T2 - T3	1.32***	-0.98*	-1.07***	0.26	0.59	0.78
T2 - T4	-0.04	-0.27	0	0.14	1.39	-0.8
T3 - T4	-1.35***	0.69	1.14***	-0.06	0.64	-0.86

Notes: Significance based on the Kruskal-Wallis test for differences in distributions
Significance levels: *p<0.10, **p<0.05, ***p<0.01



(a) Pairs 0-40

(b) Pairs 0-15

Figure 9: Cumulative stopping percentages for the second variable payment round

7.5 Appendix for ability

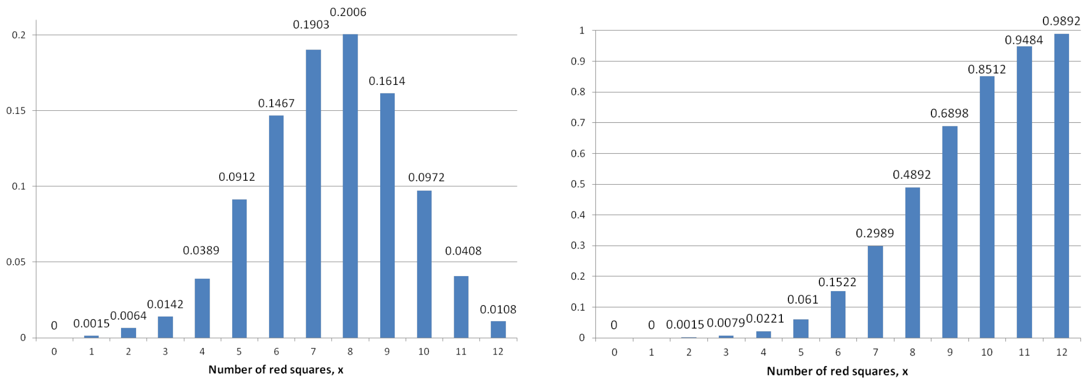
We use the last round of the test as an example of how the ability variable is calculated for each participant (grid of 7 by 8 grey squares, a maximum of twelve red squares to be found): we estimate the probability density that a random (new) participant finds exactly x red squares, with x ranging from 0 to 12. We depicted the probabilities in Figure 10a. For example, the random participant would find 12 red squares with a probability of 0.0108, 11 red squares with a probability of 0.0408, 10 red squares with a probability of 0.0972 and so on. The difficulty of finding 12 red squares for the random participant is then defined as one minus the probability of finding 12 red squares, or 0.9892. This is equivalent to the cumulative probability of finding 0 to 11 red squares. The difficulty of finding 11 red squares is then defined as one minus the probability of finding 12 and 11 red squares, or 0.9484 (see Figure 10b). Again, this is equivalent to the cumulative probability of finding 0 to 10 squares. In this way we derive a scale for difficulty to find x red squares in each round (note that not finding any red squares is not difficult at all, therefore this score is

Table 13: Median regression results for performance as dependent variable

Performance	Pairs 0-40	Pairs 0-15	Pairs 0-10	Pairs 0-5	Pairs 15-40
Model	Main effect	Interaction	Interaction	Main effect	Interaction
Effort	0.1617*** (0.0030)	0.0844** (0.0327)	0.0837*** (0.0253)	0.0954*** (0.0109)	0.1494*** (0.0346)
Ability	0.0524* (0.0298)	0.0222 (0.0605)	0.0445 (0.0517)	0.0277 (0.2031)	0.4147 (0.3685)
Difficulty	-0.1892* (0.1055)	-0.3073*** (0.1182)	-0.2586** (0.1059)	0.0013 (0.0656)	-0.7282** (0.3061)
Interaction Effort*Ability	-	0.0011 (0.0015)	0.0009 (0.0012)	-	-0.0009 (0.0018)
Constant	-0.3015 (0.9903)	2.298 (1.545)	1.480 (1.142)	-0.5344 (0.5950)	2.6797 (7.6125)
N R^2	314 0.8700	152 0.6808	124 0.5938	60 0.5583	170 0.5717

Notes: Significance levels: *p<0.10, **p<0.05, ***p<0.01, robust standard errors in brackets

always zero). We multiply the cumulative probability with the maximum number of red squares to be found in that specific round. This means that if the participant found 12 red squares, the maximum in this round, his/her score for this round equals 11.87. If the participant found 11 red squares, his/her score is equal to 0.9484 times 12 (the maximum number of that round), or 11.38. We do this for each round and sum up these scores for our ability variable.



(a) Kernel probability estimate

(b) 1-(cumulative probability of x-1)

Figure 10: Ability calculation in the last round

Table 14: Cohen's d for comparisons for high- and low-ability agents, by treatment

Performance	T1	T2	T3	T4
T1	L/H -1.88***	H 0.85**	H 2.76***	H 1.74***
T2	L -0.85**	L/H -1.00**	H -0.91**	H -0.33
T3	L -1.13***	L -0.44	L/H -0.64**	H -0.54*
T4	L -0.67*	L -0.01	L -0.38	L/H -0.57

Notes: Significance levels: *p<0.10, **p<0.05, ***p<0.01
H= high-ability, L= low-ability, column compared to line

7.6 Risk averison test

Table 15: Risk preferences test according to Holt and Laury (2002), in Euros.

Option A		Option B		Exp.PO A - Exp.PO B
Lottery	Expected payoff	Lottery	Expected payoff	
1/10 of 2 and 9/10 of 1.60	1.64	1/10 of 3.85 and 9/10 of 0.10	0.475	1.17
2/10 of 2 and 8/10 of 1.60	1.68	2/10 of 3.85 and 8/10 of 0.10	0.85	0.83
3/10 of 2 and 7/10 of 1.60	1.72	3/10 of 3.85 and 7/10 of 0.10	1.225	0.50
4/10 of 2 and 6/10 of 1.60	1.76	4/10 of 3.85 and 6/10 of 0.10	1.60	0.16
5/10 of 2 and 5/10 of 1.60	1.80	5/10 of 3.85 and 5/10 of 0.10	1.975	-0.18
6/10 of 2 and 4/10 of 1.60	1.84	6/10 of 3.85 and 4/10 of 0.10	2.35	-0.51
7/10 of 2 and 3/10 of 1.60	1.88	7/10 of 3.85 and 3/10 of 0.10	2.725	-0.85
8/10 of 2 and 2/10 of 1.60	1.92	8/10 of 3.85 and 2/10 of 0.10	3.10	-1.18
9/10 of 2 and 1/10 of 1.60	1.96	9/10 of 3.85 and 1/10 of 0.10	3.475	-1.52
10/10 of 2 and 0/10 of 1.60	2.00	10/10 of 3.85 and 0/10 of 0.10	3.85	-1.85

Table 16: CRRA estimate according to Holt and Laury (2002) and risk aversion estimate according to Brink and Rankin (2011)

Number of safe Choices[25]	Range of relative risk aversion for $U(x) = x^{1-r}/(1-r)$	Risk preference classification	Risk aversion estimate	Experimental outcomes, N=157
0-1	r < -0.95	highly risk loving	$v_1 > 1.95$	5/3.18%
2	-0.95 < r < -0.49	very risk loving	$1.95 > v_1 > 1.49$	3/1.91%
3	-0.49 < r < -0.15	risk loving	$1.49 > v_1 > 1.15$	4/2.55%
4	-0.15 < r < 0.15	risk neutral	$1.15 > v_1 > 0.85$	37 23.57
5	0.15 < r < 0.41	slightly risk averse	$0.85 > v_1 > 0.59$	29/18.47%
6	0.41 < r < 0.68	risk averse	$0.59 > v_1 > 0.32$	32/20.38%
7	0.68 < r < 0.97	very risk averse	$0.32 > v_1 > 0.03$	23/14.65%
8	0.97 < r < 1.37	highly risk averse	$0.03 > v_1 > -0.36$	6/3.82%
9-10	1.37 < r	stay in bed	$-0.36 > v_1$	18/11.46%

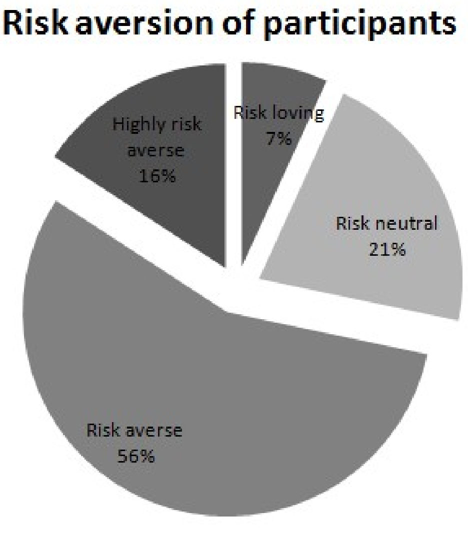


Figure 11: Percentage division of participants' risk aversion

7.7 Loss aversion test

Table 17: Loss aversion test based on Gächter et al. (2007) and experimental outcomes

Option A		Option B	Expected value Option A	Experimental outcomes, N=157	
50%	50%	100%		Rejection	Cumulative rejection
(1.00)	5.00	0	2.00	1/0.64%	1/0.64%
(2.00)	5.00	0	1.50	9/5.73%	10/6.37%
(3.00)	5.00	0	1.00	26/16.56%	36/22.93%
(4.00)	5.00	0	0.50	42/26.75%	78/49.68%
(5.00)	5.00	0	0	66/42.04%	144/91.72%
(6.00)	5.00	0	(0.50)	12/7.64%	156/99.36%
Accept all				1/0.64%	157/100%

Table 18: Loss aversion measure λ for different risk levels based on calculations of Brink and Rankin (2011)

	Upper limit for λ if Option A is chosen			
Switching Point	Risk neutral v1=v2=1.0	Reflection effect v1=v2=0.5	Always risk averse v1= 0.5, v2=1.3	Always risk loving v1=1.3, v2=0.5
BBBBBB	$\lambda > 5$	$\lambda > 2.24$	$\lambda > 2.23$	$\lambda > 8.10$
A/BBBBB	$2.5 < \lambda < 5$	$1.58 < \lambda < 2.24$	$0.91 < \lambda < 2.23$	$5.73 < \lambda < 8.10$
AA/BBBB	$\frac{5}{3} < \lambda < 2.5$	$1.29 < \lambda < 1.58$	$0.53 < \lambda < 0.91$	$4.68 < \lambda < 5.73$
AAA/BBB	$1.25 < \lambda < \frac{5}{3}$	$1.11 < \lambda < 1.29$	$0.37 < \lambda < 0.53$	$4.05 < \lambda < 4.68$
AAAA/BB	$1 < \lambda < 1.25$	$1 < \lambda < 1.11$	$0.28 < \lambda < 0.37$	$3.62 < \lambda < 4.05$
AAAAA/B	$\frac{5}{6} < \lambda < 1$	$0.91 < \lambda < 1$	$0.22 < \lambda < 0.28$	$3.31 < \lambda < 3.62$
AAAAAA	$\lambda < \frac{5}{6}$	$\lambda < 0.91$	$\lambda < 0.22$	$\lambda < 3.31$

7.8 Statistics on costs

Table 19: Percentage of agents stopping on the specified ranges in the different treatments

Payment scheme	0-10 pairs	11-20 pairs	21-30 pairs	31-40 pairs	0-40 pairs
T1	23.75%	20.00%	**7.50%**	48.75%	100%
T2	23.75%	**22.50%**	**7.50%**	46.25%	100%
T3	**85.90%**	8.97%	0%	5.13%	100%
T4	25%	19.74%	3.95%	**51.32%**	100%
Total	**39.49%**	18.79%	4.78%	37.90%	100%

Notes: Highest stopping percentage on each range in bold.
Based on the two variable payment rounds combined.