How do Incentives affect Creativity?*

Katharina Eckartz† Oliver Kirchkamp‡ Daniel Schunk§

January 9, 2013 - revision 104

Abstract

We compare performance in a word based creativity task under three incentive schemes: a flat fee, a linear payment and a tournament. Furthermore, we also compare performance under two control tasks (Raven's advanced progressive matrices or a number-adding task) with the same treatments. In all tasks we find that incentives seem to have very small effects and that differences in performance are predominantly related to individual skills.

Keywords: Creativity, Incentives, Real effort task, Experimental economics

JEL Classification: C91, J33

1 Introduction

Innovation and creativity are receiving increasing attention in research. In business, for example, a discussion emerged on how to set the conditions to achieve an optimal level of employee creativity. One potentially influential factor is the payment scheme. While it is difficult to examine this mechanism with field-data, the incentive-research in experimental and behavioural economics has mainly focused on stated effort experiments.[1] Laboratory experiments that involved real-effort tasks focused largely on production tasks, which were cognitively undemanding and did not require creativity. In this paper we attempt to close this gap and examine the impact of different payment-schemes on a creative, real effort task.

*The paper has benefited a lot from comments by participants of seminars in Jena, Mannheim, Exeter, Luxembourg, Nuremberg and Munich. We thank Jacqueline Krause, Claudia Niedlich and Severin Weingarten for assistance and support. We are grateful for financial support by the University of Jena.

†Friedrich-Schiller-University Jena, International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World. katharina.eckartz@uni-jena.de

‡Chair for Empirical and Experimental Economics, Friedrich-Schiller-University Jena, oliver@kirchkamp.de

§University of Mainz, mail@daniel-schunk.de

[1]This type of task has been used in many gift-exchange experiments; for an overview see Gächter & Fehr (2002)

Classic microeconomic labor supply theory suggests that people will provide more effort under performance-pay, irrespective of the task. This holds true also for cognitive tasks, if one regards thinking as a costly activity, as (some) economists do (discussed in Camerer & Hogarth, 1999). There are, however, several examples from the field in which incentives work counterproductively. Camerer *et al.* (1997) find that New York City Cabdrivers work less when their hourly payment is high. Dandy *et al.* (2001) find that basketball players perform better during training than during the actual game. Ariely *et al.* (2009) perform a controlled experiment in rural India where they find that performance can decrease when incentives are high.

Having said that, there are several laboratory experiments with simple real effort production tasks which find a positive impact of incentives on effort. Fahr & Irlenbusch (2000) find that their participants crack more walnuts when their wage is higher. Dickinson (1999)'s participants type more letters when their compensation depends more on their performance. van Dijk *et al.* (2001) observe that solutions for a two-variable optimisation task are better if payment is based on a tournament.

Incentives in the lab, however, are not always increasing performance. Gneezy & Rustichini (2000) find that payments for performance in an IQ-test actually decrease performance if these payments are too small. Henning-Schmidt *et al.* (2005) find no positive wage-effort relation when participants in an experiment type abstracts into a computer.

What should we expect for an experiment on creativity? Following standard labor supply theory, participants should perform better under performance-pay as compared to a flat fee. However, one factor that is completely neglected by this approach is that working on some tasks may be in itself rewarding and people might be intrinsically motivated. This may be specifically true for creative tasks.[2] Introducing incentives in such tasks might crowd out intrinsic motivation and therefore not lead to the desired result.[3]

Styhre (2008) examined in an empirical study which incentives motivate researchers. In an occupation like research where both creativity and serendipity play an important role Styhre concludes that the main factor that motivates scientists are not monetary rewards or career opportunities but the excitement of discovering an unknown domain. Due to the dependence on serendipity, a researcher's motivation to be creative decreases when being under constant pressure to deliver outputs or to fulfil increasing demands.

A related experimental study, focusing on innovation, is Ederer & Manso (2008). They study behaviour of participants in an experiment who operate an

[2]There is a large body of experimental psychological research on creativity among others by Amabile and her co-authors as well as by Sternberg and co-authors. This research focuses however, when looking at rewards, mainly on reward- versus non-reward scenarios. From this research it seems that the effects of rewards on creativity depend among others on the task type, the initial levels of intrinsic motivation and the salience of the extrinsic reward. While Amabile notes that it is easier to find laboratory conditions which decrease creative performance, she also identifies conditions under which intrinsic motivation and extrinsic rewards can be additive.

[3]Motivational crowd out is also discussed in the experimental literature in the context of imposing minimal-effort levels and monitoring (among others Falk & Kosfeld (2006) and Ziegelmeyer *et al.* (2011).)

artificial lemonade stand which profits depended on the chosen location (exploration) and the product-characteristics (exploitation), while the optimal product-mix was different for the various locations. Ederer & Manso then compared different wage-schemes: fixed wage, performance-based-pay and an "exploration contract". The latter is a partly-performance-based-pay contract: the payoffs depended on the profits during the second half of the experiment. This gave subjects the possibility to first explore and by that included a "tolerance for early failure". The authors find that this exploration contract performs better than standard fixed wage or performance-based-pay contracts. In contrast to Ederer & Manso who use an exploration task we will put our focus on creativity.

We will describe our experiment in section 2, report our results in section 3 and conclude in section 4.

2 Experiment

2.1 Tasks

In this study we investigate in a *within-subject* design how participants perform under different incentive schemes in a task, which requires not only cognitive efforts but also creative thinking. We run a pure cognitive effort task as a control.

Finding a task for the experiment that requires creative thinking did not turn out to be easy. Requirements were that the quality of the solution is easy to assess and that the task remains interesting when it is repeated. Specific problems like insight problems (e.g. Schooler *et al.* (1993)) or packing quarters into a box, a task which has been used by Ariely *et al.* (2009), are easy to assess but can be used for each participant only once. After a single round of a treatment participants have understood the problem and will, with or without incentives, quickly be able to apply the solution again.[4]

Open tasks like "painting a creative picture" might remain interesting even after several pictures, but it would be hard for the experimenter to judge the quality of the solutions that are produced in the laboratory. The "standard" procedures to use experts (Amabile, 1996), one or more researchers, a larger group of students, or a web based tool (Girotra *et al.*, 2009), to assess the quality of submissions would all take too much time in a repeated laboratory experiment. Here we will use tasks that can be quickly and mechanically rated by the computer.

Word task: In our study we use a word task[5] as our creative thinking task: participants are presented with an alphabetically ordered letterset, consisting of 12 letters, e.g. `accdeeeginst`. Participants have to construct as many words as they can within 5 minutes. Rewards were more than proportionally increasing with the length of the created word (see section 2.2 for a detailed overview). Table 1 gives some examples of words that can be constructed with these letters as well

[4]For insight problems, like the well-known candle problem (Duncker & Lees, 1945), participants that came across the problem before will immediately know the solution.

[5]This task is partially inspired by word games like Scrabble, partially by a task that Crosetto (2010) used to simulate sequential innovation in the lab.

Table 1 Example: words that can be constructed with `accdeeeginst`

`a`	1 point
`ac`	1+2=3 points
`and`	1+2+3=6 points
⋮	
`teasing`	1+2+3+4+5+6+7=28 points
`accidents`	1+2+3+4+5+6+7+8+9=45 points

as the resulting points.[6] Appendix A.1 shows all English words that a participant could find for the above letterset. Appendix A.2 shows all German words for a similar letterset.

We find that such a "word task" has many aspects of a creative task and that it mimics quite well a creative innovation. Whenever an inventor invents something, an idea is generated and tested against the inventor's model of nature. The Eureka! moment is the realisation that the idea, often a composition of several simpler principles, passes this test. Similarly in our word task participants have to generate words (not entire ideas, though) and test these words against a simple model of nature, here a dictionary. We concede that the pure exploration aspect of research is not captured by our task. E.g. a developer of a drug who has no idea at all what type of drug might work and who is exploring the range of possible drugs in an unsystematic way is not captured by our model. We suspect, however, that many inventors have a quite good model of the world which is relevant for them, that they search in a structured way for solutions, and that a main and creative ingredient of invention is the realisation that ingredients A, B, and C can be combined in a clever way in order to create D. Patented inventions like the suspension bridge, the commutator type electric motor, the Yale lock, the sewing machine, the milking machine, the safety pin, the mouse trap, barbed wire, the ball-point pen, the zipper, the adjustable wrench, disk brakes, the supermarket, frozen food, the banana protective device, the ice cream bar, the monopoly game, the Lego brick, or the bathing suit are all obvious once one "gets the idea". In all these cases getting the idea meant putting the underlying principles together.

When designing the lettersets we were aiming at using lettersets which are very similar to each other on a number of potentially relevant dimensions. To create these lettersets we first randomly build 100 000 different lettersets and then determined which words could be constructed out of each set by comparing possible words with the German isoword-list (Knutzen, 1999). This list contains 294897 different words, including forms of words, names, abbreviations, but no swearwords. For all our 100 000 different lettersets we calculated the number of points which could potentially be constructed with each of the lettersets and finally chose the lettersets which were similar in three dimensions: the number of points that could be earned, the number of words that could be created and the similarity among the words.[7] The resulting eight lettersets are displayed in table 2.

After a pilot in which we used all 8 lettersets, we dropped the 2 best- and the

[6]Since we ran the experiment in Germany, we used German words.

[7]We used the fstrcmp form GNU Gettext 0.17 to calculate for each word the similarity to the most similar word in the set.

Table 2 Lettersets

letters	points	words	similarity within
aceehhinrssä	5501	323	0.886879
cdehhlorsstt	5445	323	0.886458
aehklllprstt	5386	326	0.886948
aeeeggllmnru	5430	323	0.886883
deehhimnnprt	5449	321	0.886626
aaeehhiknstt	5503	329	0.886679
cdeeillrsstw	5427	327	0.887130
deegilmnnpuw	5405	322	0.887139

Table 3 Raven's matrices

Subset	matrix number
1	1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, 34
2	2, 5, 8, 11, 14, 17, 20, 23, 26, 29, 32, 25
3	3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36

2 worst-scoring ones. Table 4 shows which lettersets were used in the final experiment. During the experiment participants received a feedback after each word-submission on whether the word they entered was accepted, entered wrongly or had been entered before. All correct words were shown as a list on the screen. Participants were not informed about how many points they had.

Control tasks: The control tasks differ between the two experimental series. In the first and main experimental series this control task was an IQ-task. In the second experimental series this control task was a number-adding task.

IQ task: The IQ-task was based on an intelligence test, Raven's advanced progressive matrices, set II (see Raven *et al.*, 1998). Raven's matrices are designed to measure eductive ability: the ability to make sense of complex facts and reproductive ability, i.e. the ability to store and reproduce information. These two components had been identified by Spearman (1923, 1927) as being the two main components of general cognitive ability. The version of Raven's matrices we used in this experiment was the one designed for subjects with high ability. The set consists of 36 matrices which are increasingly difficult. Since we also wanted to use a within participants design for the intelligence task we split this set into three subsets: the matrices were alternatingly distributed on the three subsets to ensure that the three subsets are of approximately the same difficulty (see table 3).

Number-adding task: In a second experimental series we replace the IQ-task with a number-adding task, similar to the one used by Niederle & Vesterlund (2007): participants had to add for five minutes five two-digit numbers.[8] Participants were allowed to use scratch-paper for their calculations. Moreover, after

[8]E.g.: 12 + 73 + 05 + 56 + 60. The numbers were drawn randomly. The same numbers were presented to all participants in the same order.

each summation, participants received feedback on whether their solution was correct.

While the performance in the IQ-task may depend mainly on ability, the number adding task depends clearly, as also Niederle & Vesterlund note, on skill and effort. In our opinion the skill component in this task should be less pronounced than in the IQ-tasks, which may lead to more response to the experimental treatments than the pure IQ-task.

Questionnaire: At the end of the experiment participants answered a questionnaire including questions on participants' task-interest for the two tasks as well as how much they enjoyed working on the two tasks. Moreover we collected demographics and language skills. Since preferences for payment-schemes might be related to the participants' risk-preferences, we elicited those at the end of the experiment using the risk-question of Dohmen *et al.* (2011).[9]

2.2 Treatments

We are interested in differences in participants' performance under different payment schemes in a given time. In this experiment we compared three different schemes: a *flat fee* regime, a *linear* payment regime and a *tournament*. All parameters were calibrated such that the expected payment for the experiment that lasted for approximately one hour was about 10€. This is considerably more than the average hourly wage of a student assistant at the University of Jena. In this experiment the focus is on how different treatments influence the effort provided by the experimental subjects.[10] Higher effort will result in higher output only if the match between the task and the subject is good enough as, among others, Camerer & Hogarth (1999) argue. We believe that our mainly student subject pool satisfies this criterion.

The experiment consisted of seven stages, each lasting five minutes. In each treatment participants always started with the creativity task and afterwards solved the control task with the same incentive scheme. We varied the sequence of treatments to rule out order effects. No feedback was given during the experiment. Table 4 provides an overview.

During the experiment participants received points for correct solutions. At the end of the experiment one of the seven stages was randomly selected for payment.[11] The respective number of points was converted to Euros with an exchange rate of 1 point = 0.04€. In the flat scheme participants received 250 points (=10€) irrespective of their performance. In all three conditions the instructions asked the participants to build as many and as long words as possible. In the two performance-pay conditions, we rewarded the obvious increasing difficulty to construct long words with more than proportionally more points. More

[9] Dohmen *et al.* (2011) included the question in the 2004 wave of the German Socio Economic Panel. They found this question to be correlated with real risk-taking behavior while a lottery choice did not predict real risk-taking behavior as well as the simple question.

[10] In contrast to other studies who focus on the provided working-time.

[11] We do this to prevent participants from hedging between stages.

Table 4 Experimental design

stage	letterset / matrix subset	treatment*
1	`aceehhinrssä`	treatment 1
2	subset 1 (1, 3, 7, ...)	treatment 1
3	`aeeeggllmnru`	treatment 2
4	subset 2 (2, 4, 9, ...)	treatment 2
5	`deehhimnnprt`	treatment 3
6	subset 3 (3, 5, 10, ...)	treatment 3
7	`deegilmnnpuw`	self-selection
	Questionnaire	

* The treatment order was alternating for different individuals, i.e. for some individuals treatment 2 had flat incentives, for other individuals treatment 2 was, e.g., linear incentives.

specifically, points were awarded such that participants received for every correct word they produced 1 point for the first letter, 2 points for the second, 3 for the third and so on. This means that a word with 5 letters was awarded with 5+4+3+2+1 = 15 points (see table1). In the control task the number of points per correct solution was constant: every correctly solved IQ-task was awarded with 60 points while every correctly solved number-adding task was awarded similarly with 25 points.[12] In the tournament participants were matched with three other participants and the number of acquired points was compared for the respective task. A winning participant was awarded 25€ (if that condition was chosen for payment) and a losing participant was compensated with 5€. The size of these prizes was chosen such that the winning prize was substantially higher than the size of the losing prize. We decided not to use a "winner-takes-it-all" design in the tournament but to also compensate the losing participants with a small prize to give participants a small compensation for showing up and putting effort into the experiment.[13]

The last stage of the experiment was a self-selection stage. Participants could chose which of the previously experienced payment-schemes they preferred for the subsequent word-production-task. If they opted for the tournament condition their performance was compared to the previous performance of their matching group members in the first tournament-condition. This was done to avoid confounding preferences for a payment-scheme with beliefs about who might enter a tournament.[14] We included the self-selection stage as this allows us to investigate several questions: who selects which payment-scheme, do we find differences in

[12]The piece-rate in the IQ-task and the creativity task were based on our pilot experiment, the piece-rate in the number-adding task was based on the average number of correct solutions in Niederle & Vesterlund (2007).

[13]If in the end a tournament stage was chosen for payment, then points were compared within a group of four participants who were all facing the same sequence of treatments. Eventual ties were broken randomly and automatically. Otherwise, participants were working independently throughout the experiment. They received no information about the identities or the results of other participants.

[14]See, e.g., Niederle & Vesterlund (2010).

performance following self-selection and if so, whether this represents sorting. Niederle & Vesterlund (2007) find gender-differences in the choice of the preferred payment scheme in their number-adding-task: having to chose between a tournament and a linear payment scheme, 73% of the men and less than half as many women (35%) chose the tournament. In a stated-effort experiment Eriksson *et al.* (2009) look, among others, on the impact of risk preferences and find that risk-averse subjects are less likely to enter tournaments.

2.3 Conducting the experiment

The main experiment was conducted in November and December 2010 in the laboratory of the Friedrich-Schiller-University in Jena. Three additional sessions were run in June 2011.[15] In total the experiment was run in 13 sessions, each having between 14 and 18 participants. In total 216 participants took part in the experiment, of which 50 participated in the second experimental series. Since the experiment contains a tournament treatment, we deliberately invited an equal number of men and women for every session so that potential group-composition effects concerning gender are kept as similar as possible. Small differences are due to non-show-ups. Overall, however, the gender composition was balanced within and across sessions (see the left graph in Figure 10 in the appendix). Before the experiment started participants were waiting in the corridor, so they were aware of the composition of the experimental group.[16]

Of the 216 participants, 198 were undergraduate students of a broad variety of fields of study. The average age of all participants was 23.7. Participants were recruited on-line using ORSEE (Greiner, 2004).

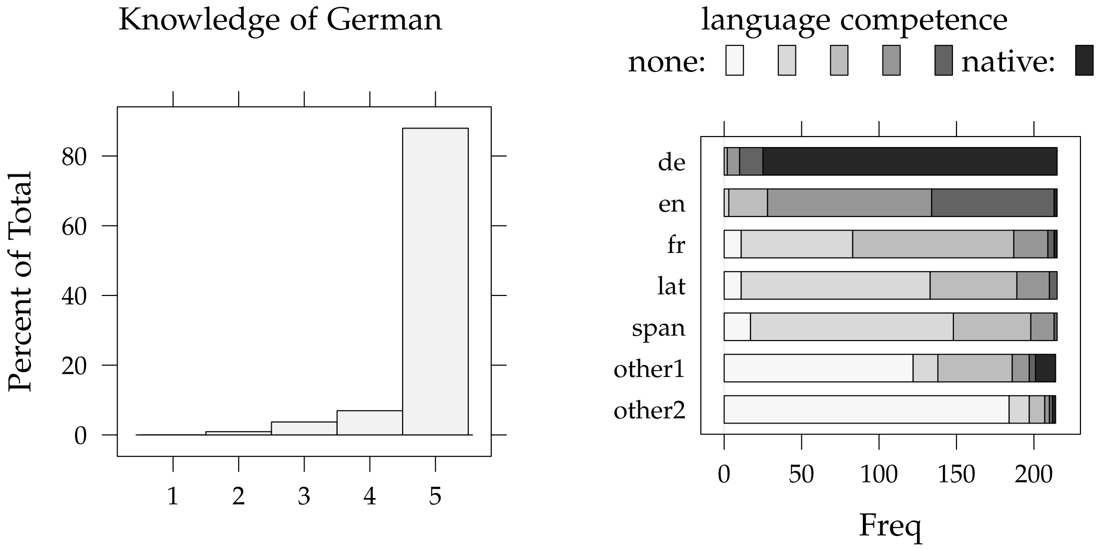
At the end of the experiment the computer chose one of the 7 stages for payment. The payment-procedure was as follows: we first distributed the receipts and then participants exchanged signed receipts for an envelope with their payment. All sessions lasted for about one hour. The average payment amounted to 10.31€.

The language of instruction was German and participants were informed in the invitation to the experiment that knowledge of German at the level of a native speaker was necessary to be able to participate in the experiment. They also knew that they had to pass a short language-test previous to the experiment (unless they had already passed this test during an earlier experiment). Only participants who had passed this test were admitted to the experiment. The experiment was programmed browser-based using PHP in combination with a MySQL database and an Apache server. All entered words were spell-checked and only words which were spelled correctly were accepted. The browser-settings were set such that the participants saw the experiment on a full screen, just like in any other experiment. The use of the keyboard was restricted such that participants neither had the possibility of moving back- or forwards in the experiment nor could they leave the full-screen mode.

[15]The IQ-task was used as a control-task in the experiments in 2010 while the number-adding task was used in the three experiments in June 2011.

[16]Nobody in the experiment was aware of the identity or gender of their matching group members.

Figure 1 Language competence



3 Results

Questionnaire data Since the creative task requires very good knowledge of German, we not only required the participants to pass a short language test, but in addition participants also rated their language skills on a scale from 1 to 5, where 1 represented no knowledge of the language and 5 represented knowledge at the level of mother tongue. The average self-reported language-knowledge of German was 4.8 on a scale from 1 to 5. In addition information about the knowledge of other languages was also collected. The distribution of the language competence for German and other languages is shown in figure 1.

We also collected information about the participants' hobbies, in particular whether they enjoy reading, discussing, solving crossword puzzles, playing scrabble, being creative and solving logic-puzzles. While the first four obviously are related to the lexis of the participants and their joy of doing word-related task, the last one is collected to have a control variable which might be related to solving Raven's Matrices (Figure 2). To assess participants' interest for creative tasks, we included in addition to the question about creativity as a hobby also a questionnaire on self-reported creative potential in the post-experimental questionnaire (DiLiello & Houghton, 2008). An overview is given in Figure 3.

Risk-preferences were elicited with the risk-question (Dohmen *et al.*, 2011) which is a 11-point scale, reaching from 0 (being very risk-averse) to 10 (being very risk-loving). The distribution is shown in Figure 3.

Aggregate performance To assess whether we rely on different or rather similar skills with the different tasks we show 95% confidence intervals (based on an ABC bootstrap) for correlations of the performance for the different tasks in figure 4. We see that participants who perform well in one stage in the word task also perform well in the next stage. Similarly, performance within each of the control tasks is correlated. However, correlation of performance in the word task with

Figure 2 Hobbys and interest in languages

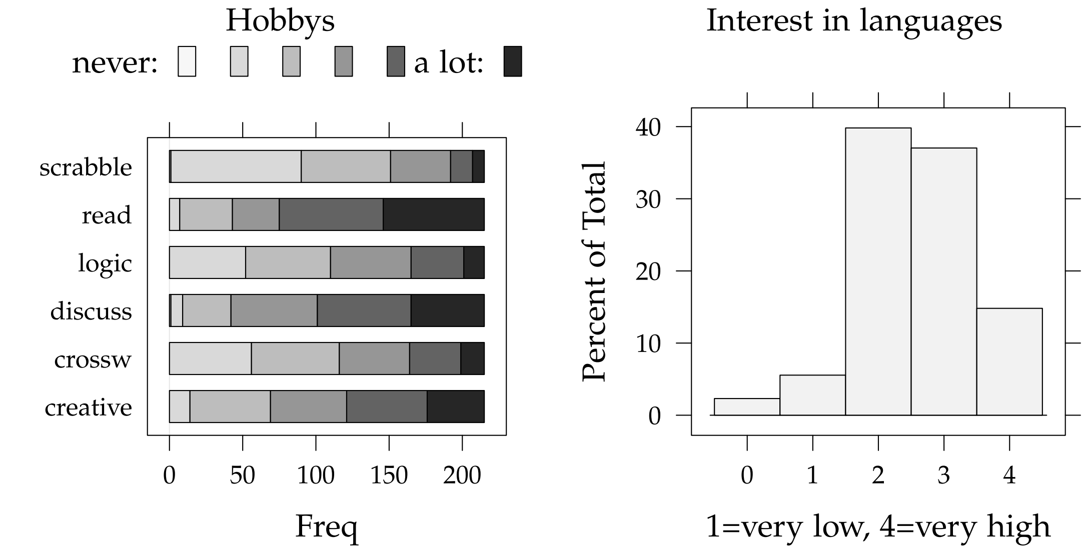


Figure 3 Creativity and attitude toward risk

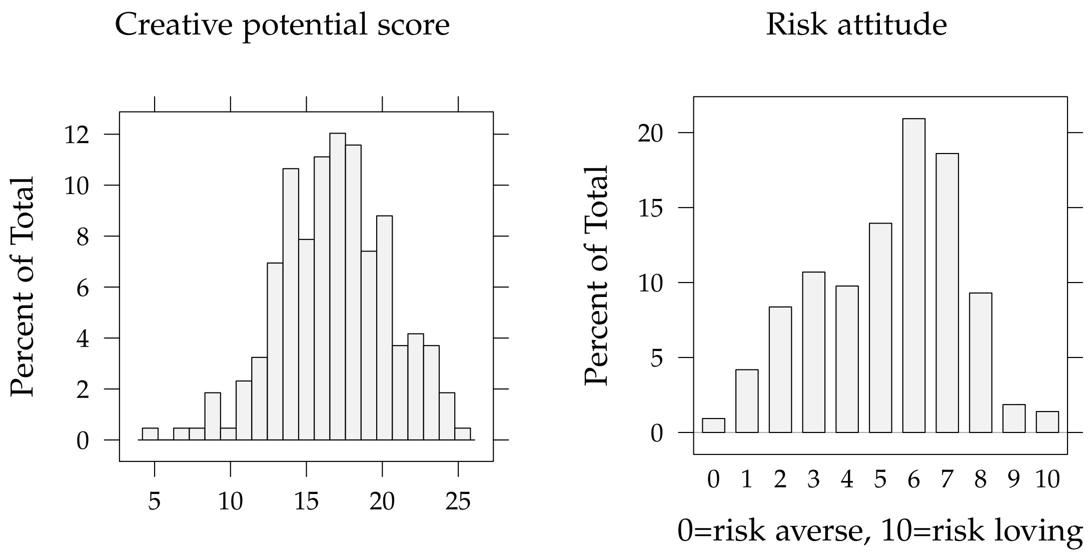
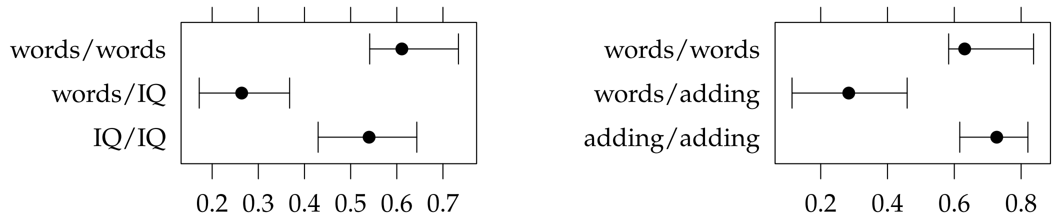


Figure 4 Correlation of performance among different tasks



The segments show 95%-confidence intervals (based on ABC bootstraps). The left graph shows data from the treatment with words and IQ-task, the right graph shows data from the treatment with words and number-adding.

Table 5 Average contribution to R^2 in %

	words	IQ	number-adding	length	distance
subject	69.17	67.27	81.07	57.02	5.34
stage	6.80	2.74	0.94	0.19	1.65
incentive	0.38	0.22	0.05	0.77	0.03

Contributions for words, IQ, and number-adding are based on equation (1), for length on equation (3) and for distance on equation (4).

performance in the control task is much lower. Though still positive, we can say that words and both control tasks seem to depend on quite different skills.

In a next step we want to find out whether incentives have a substantial influence on performance. To do this we compare the effect of incentives on performance with the effect of individual heterogeneity (a dummy for the participant) and possible learning effects during the experiment (measured as a dummy for the stage in the experiment). We estimate the following equation:

$$\text{Performance} = \sum_{\text{Subjects}} \beta_{\text{subj.}} d_{\text{subj.}} + \sum_{\text{Stages}} \gamma_{\text{st.}} d_{\text{st.}} + \sum_{\text{Incentives}} d_{\text{inc.}} \delta_{\text{inc.}} + \epsilon_i \qquad (1)$$

The average contribution of the regressors to the R^2 (following Lindeman *et al.*, 1980, p. 119ff) is shown in table 5. We find that for all treatments, words, IQ, and number-adding, the impact of the incentive scheme on performance is very small compared to individual heterogeneity (measured as "subject") or even compared to learning (measured as the "stage").

To assess the magnitude of the effect in absolute terms we estimate the following mixed effects equation:

$$\text{Performance} = \beta_0 + \sum_{\text{Incent.}} \beta_{\text{inc.}} \cdot d_{\text{inc.}} + \epsilon_{\text{stage}} + \epsilon_{\text{subj.}} + \epsilon_{\text{subj.},t} \qquad (2)$$

In this equation the incentive scheme *flat* is the baseline, $\epsilon_{\text{st.}}$ is a random effect for the stage, $\epsilon_{\text{subj.}}$ is a random effect for each individual participant and $\epsilon_{\text{subj.},t}$ is

Table 6 Estimation results for equation (2) for words

	β	σ	t	p value	95% conf	interval
(Intercept)	256	71.6	3.57	0.0004	115	396
linear	14.7	8.69	1.69	0.0920	-2.41	31.8
tournament	15.7	8.81	1.78	0.0755	-1.62	33

Table 7 Estimation results for equation (2) for IQ

	β	σ	t	p value	95% conf	interval
(Intercept)	6.16	0.583	10.6	0.0000	5.01	7.31
linear	0.14	0.184	0.764	0.4456	-0.221	0.502
tournament	0.231	0.183	1.26	0.2083	-0.129	0.592

the residual. Estimation results for words, IQ, and number-adding are shown in Tables 6, 7, and 8, respectively. While the treatment effects are small for all tasks, they are not significant for both control tasks and only significant at a 10%-level in the word task.

Complexity and originality In reality, firms might not mainly be interested in the number of creative answers to one question, but rather in having one single high-quality solution. Above we have seen that incentives do not change very much the overall productivity of participants in our experiment. It might still be that incentives affect the quality. In the context of our word task we might suspect that incentives have an effect on complexity or originality.

E.g. with the letterset `accdeeeginst` a participant could produce many short and simple words like `a`, `i`, `dan`, or `ian`. A participant could also think harder and produce longer and more complex words like `accidents` or `deceasing`. Since `accidents` has a value of 45 points and `dan` has only a value of 6 points some participants might find it more profitable to spend more time looking for longer words.

Another relevant dimension might be originality of the product. Participants might resort to a sequence of rather similar items like `cease`, `ceased`, and `ceasing` or they might turn out to be more original and create words that are more diverse like `denis`, `ideas`, `stance`, etc. We measure dissimilarity as the Jaro-Winkler Distance of successive words (Jaro, 1989, Winkler, 1990).

Table 8 Estimation results for equation (2) for number-adding

	β	σ	t	p value	95% conf	interval
(Intercept)	10	2.72	3.68	0.0004	4.61	15.4
linear	0.085	0.621	0.137	0.8915	-1.15	1.32
tournament	0.22	0.647	0.34	0.7347	-1.06	1.5

Table 9 Determinants of word length, equation (5)

	β	σ	t	p value	95% conf	interval
(Intercept)	4.34	0.0487	89.1	0.0000	4.24	4.43
linear	0.0864	0.0385	2.24	0.0255	0.0106	0.162
tournament	0.0186	0.0385	0.482	0.6301	-0.0572	0.0943

Table 10 Determinants of distance among words, equation (6)

	β	σ	t	p value	95% conf	interval
(Intercept)	0.611	0.0504	12.1	0.0000	0.513	0.71
linear	0.00586	0.00509	1.15	0.2493	-0.00411	0.0158
tournament	0.0061	0.00503	1.21	0.2250	-0.00375	0.016

We estimate the following two equations:

$$\text{Length} = \sum_{\text{Subjects}} \beta_{\text{subj.}} d_{\text{subj.}} + \sum_{\text{Stages}} \gamma_{\text{st.}} d_{\text{st.}} + \sum_{\text{Incentives}} d_{\text{inc.}} \delta_{\text{inc.}} + \epsilon_i \quad (3)$$

$$\text{Distance} = \sum_{\text{Subjects}} \beta_{\text{subj.}} d_{\text{subj.}} + \sum_{\text{Stages}} \gamma_{\text{st.}} d_{\text{st.}} + \sum_{\text{Incentives}} d_{\text{inc.}} \delta_{\text{inc.}} + \epsilon_i \quad (4)$$

Table 5 also shows the average contribution of our regressors to the R^2 (Lindeman *et al.*, 1980, p. 119ff) for equations (3) and (4). For comparison the table also shows the contributions to the equation for performance, equation (1). Similar to productivity (see above) also the (aggregate) impact of incentives on the type of the product, either measured as size (word length) or diversity (Jaro-Winkler distance) is very small.

To measure the absolute magnitude of the effect we also estimate the following mixed effects model:

$$\text{Length} = \beta_0 + \sum_{\text{Incentives}} \beta_{\text{inc.}} \cdot d_{\text{inc.}} + \epsilon_{\text{stage}} + \epsilon_{\text{subj.}} + \epsilon_{\text{subj.},t} \quad (5)$$

Estimation results are shown in table 9. We see that incentives do have a positive impact on word length, however, only the effect of linear incentives is significant.

To measure the absolute impact of incentives on originality we estimate the following mixed effects equation:

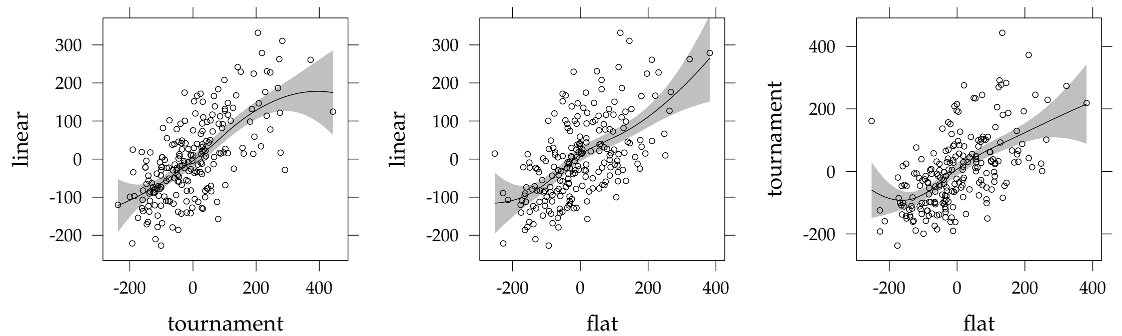
$$\text{Distance} = \beta_0 + \sum_{\text{Incentives}} \beta_{\text{inc.}} \cdot d_{\text{inc.}} + \epsilon_{\text{stage}} + \epsilon_{\text{subj.}} + \epsilon_{\text{subj.},t} \quad (6)$$

Estimation results are shown in Table 10. The impact of incentives is positive, but small and not significant.

Individual heterogeneity Although aggregate reaction to incentives is low (as we have seen above), sensitivity to incentives varies from individual to individual. To measure individual sensitivity to incentives we estimate the following regression

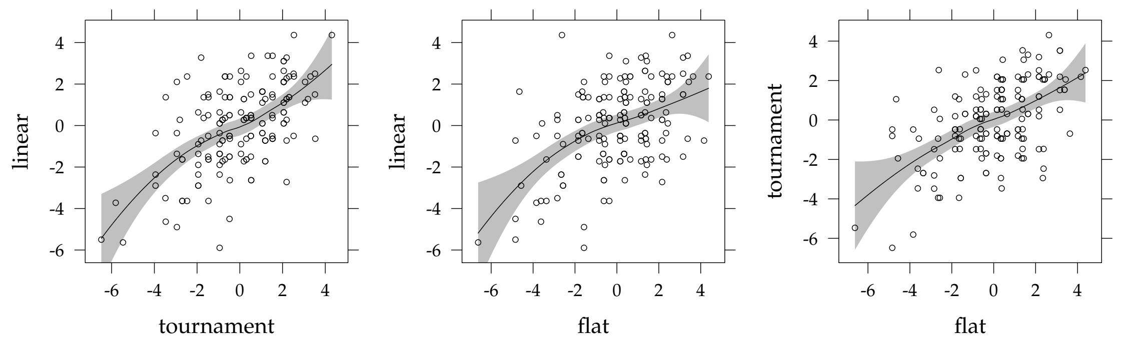
$$\text{Performance} = \sum_{\text{Incentives}} \left(\beta_{\text{inc.}} \cdot d_{\text{inc.}}\right) + \sum_{\text{Histories}} \beta_{\text{hist.}} \cdot d_{\text{hist.}} + \epsilon_{\text{subj.,inc.}} \quad (7)$$

Figure 5 Individual sensitivity to incentives for the word task, equation (7)



$r = 0.670$, $[0.588, 0.737]$ $r = 0.598$, $[0.509, 0.678]$ $r = 0.580$, $[0.475, 0.658]$
(95% confidence intervals for correlations are based on an ABC bootstrap)

Figure 6 Individual sensitivity to incentives for the IQ task, equation (7)



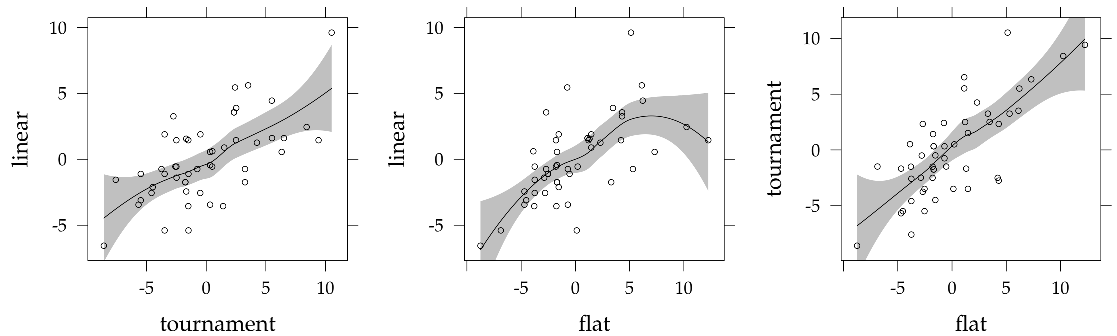
$r = 0.574$, $[0.441, 0.684]$ $r = 0.475$, $[0.320, 0.605]$ $r = 0.554$, $[0.420, 0.665]$
(95% confidence intervals for correlations are based on an ABC bootstrap)

where $\epsilon_{\text{subj.,inc.}}$ measures the (remaining) individual component of sensitivity. Figure 5 shows the joint distribution of $\epsilon_{\text{subj.,inc.}}$ for the different incentive schemes for the word task. We see that residual performance $\epsilon_{\text{subj.,inc.}}$ for the different incentives is always positively correlated. Participants who perform relatively well under one incentive mechanism also perform well under the other.

We find the same effect for IQ (Figure 6) and for number-adding (Figure 7). In all cases performance is positively (and significantly so) correlated.

Self-selection: In the last stage of the experiment subjects have the choice to select the payment scheme for another round of the word task. We see from Figure 8 that flat incentives are slightly more popular (40.74%), in particular for females (45.45%), while males seem to be relatively more interested in linear incentives (35.85%). Tournaments seem to be the least favoured choice (chosen by 28.3% of males and 26.36% of females). In contrast to Niederle & Vesterlund (2007), who find that significantly more male than female participants chose the tournament over a linear payment-scheme, we do not observe gender-differences in the likelihood of selecting the tournament. We can, however, not say where this difference

Figure 7 Individual sensitivity to incentives for the numbers task, equation (7)



$r = 0.639$, $[0.478, 0.790]$ $r = 0.627$, $[0.445, 0.766]$ $r = 0.760$, $[0.593, 0.860]$

(95% confidence intervals for correlations are based on an ABC bootstrap)

Table 11 Multinomial logit for treatment selection in the final stage, equation 8

	β	σ	t	p value	95% conf	interval
linear:(intercept)	-2.94	0.731	-4.02	0.0001	-4.38	-1.51
tournament:(intercept)	-2.4	0.723	-3.32	0.0009	-3.82	-0.982
linear:points	0.00218	0.000621	3.51	0.0005	0.000961	0.00339
tournament:points	0.00157	0.00063	2.48	0.0130	0.000331	0.0028
linear:risk	0.239	0.0838	2.85	0.0044	0.0743	0.403
tournament:risk	0.19	0.0839	2.27	0.0233	0.0258	0.355
linear:female	-0.547	0.348	-1.57	0.1156	-1.23	0.135
tournament:female	-0.321	0.352	-0.912	0.3617	-1.01	0.369

"flat" is the reference treatment. Effects are shown for the treatments "linear" and "tournament".

in observations stems from: whether it is task-specific[17], follows from differences in the experimental design[18] or whether it is subject-pool specific.

One potential determining factor for the self-selection are subjects' risk preferences. The left graph in Figure 9 shows that the likelihood of chosing the flat payment scheme decreases with more risk-loving risk preferences. Subjects' choice is also likely to be influenced by their ability. Here we interpret the number of previously acquired points in the word task as a measure of task-related ability. Looking at the right graph of Figure 9 it seems that the likelihood to switch from flat to either the linear or the tournament based payment increases with higher performance in the previous word-creation-stages. To confirm what we see in the figures we estimate the following multinomial logit model:

$$\frac{\log \Pr(\text{treatment})}{\log \Pr(\text{flat})} = \beta_{(\text{intercept})} + \beta_{\text{points}} \cdot \text{points} + \beta_{\text{risk}} \cdot \text{risk} + \beta_{\text{female}} \cdot \text{female} \quad (8)$$

[17]In Niederle & Vesterlund (2007) participants chose their payment-mechanism for the number-adding taks while in our experiment partipants made this choice for the word task.

[18]Our tournament design differed from the one used in Niederle & Vesterlund (2007) in that we offered subjects a larger choice-set and in that the tournament design was slightly different. Niederle & Vesterlund implemented a tournament in which the winner was compensated proportionally to the number of solved tasks, the loser received nothing.

Figure 8 Self selection into treatments

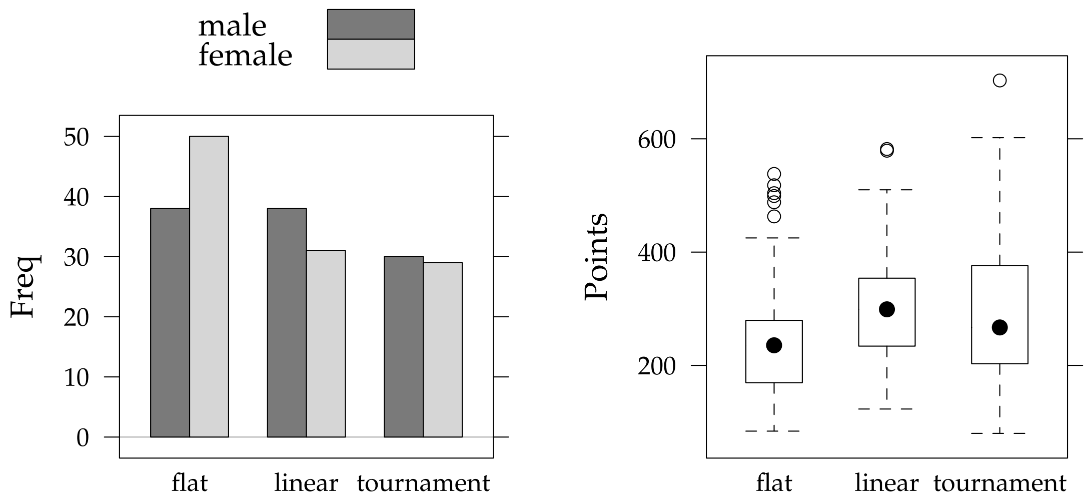
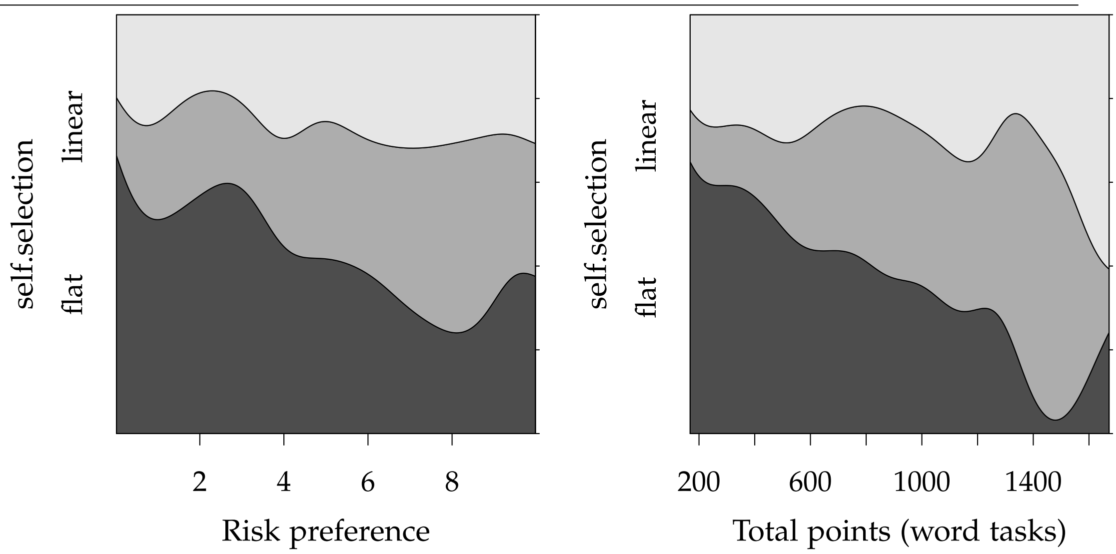


Figure 9 Self selection into treatments



We take "flat" as the reference treatment, i.e. "treatment" is either "linear" or "tournament". "Points" is the sum of points obtained in the previous three rounds of the word task (as in Figure 9). "Risk" is the risk measure according to Dohmen *et al.* (2011). Estimation results are reported in Table 11. We see that a good performance in the previous rounds makes it more likely to choose an incentivised treatment. This effect is significant for both linear and tournament, with no significant difference between the two ($p = 0.3118$). Also, more risk loving participants are more likely to select into the incentivised treatments. Again, there is no difference between the effect of risk to select into the linear incentive or the tournament ($p = 0.5878$). Finally, there is no significant effect of gender to select in one of the incentivised treatments.

Performance in the self-selection stage, as shown in the box-plot in the right part of Figure 8, differs between the selected treatments: it seems that participants who selected the flat fee obtained fewer points than those who chose performance-pay. The seemingly higher performance under performance-pay can be interpreted as sorting since the likelihood into select into a performance based payment-schemes (linear or tournament) increases with the ability (measured as total points). Concluding, with more risk-loving preferences or more points in the previous stages, people switch from flat fee to a performance based payment scheme.

4 Conclusion

Using three different tasks, one based on creativity, one on intelligence, and one adding numbers, we have seen that performance depends almost entirely on individual characteristics of participants and can, on the aggregate level, hardly be influenced through incentives. Neither on the aggregate nor on the individual level do we find effects of incentives on performance. We also do not find an effect of incentives on the similarity or complexity of generated words in the creativity task. In the self-selection stage we find no relation between gender and the choice of the tournament. In our experiment it seems that the more able and the more risk-loving people are, the more likely they are to chose an incentivised payment-scheme in contrast to a flat fee. Also we observe higher output in the performance-pay treatment after self-selection.

Given the mixed evidence from many other experiments with real efforts we should be careful in generalising our observations. Still, our results seem to support the view that effects of incentives for a range of tasks, from creative tasks to repetitive calculations, are, if at all, very small. Individual characteristics explain for all tasks more than 60% of the observed variance in the performance. The presence or absence of different incentive schemes explain for all tasks in this experiment less than 1% of the variance.

To us it is in particular striking that we do not observe effects of incentive-schemes in the control tasks. In a follow-up study we check potential factors that might explain this result. In particular we analyse whether taks-enjoyment or the availability of opportunity costs contribute to the result. We find that making tasks more difficult or less interesting does not change our results. With the introduction of opportunity costs, however, we observe differences of incentive-schemes on subjects performance.

References

Amabile, T. (1996). *Creativity in context*, Westview press.

Ariely, D., Gneezy, U., Loewenstein, G. & Mazar, N. (2009). Large stakes and big mistakes. *Review of Economic Studies*, *76*, pp. 451–469.

Camerer, C., Babcock, L., Loewenstein, G. & Thaler, R. (1997). Labor supply of new york city cabdrivers: One day at a time. *The Quarterly Journal of Economics*, *112*, pp. 407–441.

Camerer, C.F. & Hogarth, R.M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*(1-3), pp. 7–42.

Crosetto, P. (2010). *To patent or not to patent: A pilot experiment on incentives to copyright in a sequential innovation setting*, Departemental Working Papers 2010-05, Department of Economics, Business and Statistics at Università degli Studi di Milano.

Dandy, J., Brewer, N. & Tottman, R. (2001). Self-consciousness and performance decrements within a sporting context. *Journal of Social Psychology*, *141*, p. 150–152.

Dickinson, D. (1999). An experimental examination of labor supply and work intensities. *Journal of Labor Economics*, *17*(4), pp. 638–670.

DiLiello, T.C. & Houghton, J.D. (2008). Creative potential and practised creativity: Identifying untapped creativity in organizations. *Creativity and Innovation Management*, *17*, pp. 37–46.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J. & Wagner, G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*.

Duncker, K. & Lees, L. (1945). On problem-solving. *Psychological monographs*, *58*(5), p. i.

Ederer, G. & Manso, F. (2008). Is pay-for-performance detrimental to innovation?, Working Paper.

Eriksson, T., Teyssier, S. & Villeval, M. (2009). Self-selection and the efficiency of tournaments. *Economic Inquiry*, *47*(3), pp. 530–548.

Fahr, R. & Irlenbusch, B. (2000). Fairness as a constraint on trust in reciprocity: Earned property rights in a reciprocal exchange experiment. *Economic Letters*, *66*, pp. 275–282.

Falk, A. & Kosfeld, M. (2006). The hidden costs of control. *The American economic review*, pp. 1611–1630.

Girotra, K., Terwiesch, C. & Ulrich, K.T. (2009). *Idea generation and the quality of the best idea*, Research Paper 2009/65/TOM, INSEAD Business School.

Gneezy, U. & Rustichini, A. (2000). Pay enough or don't pay at all. *Quarterly Journal of Economics*, *115*(3), pp. 791–810.

Greiner, B. (2004). An online recruitment system for economic experiments, in (K. Kremer & V. Macho, eds.), *Forschung und wissenschaftliches Rechnen*, vol. 63, of *GWDG Bericht*, pp. 79–93, Göttingen: Ges. für Wiss. Datenverarbeitung.

Gächter, S. & Fehr, E. (2002). Fairness in the labour market, in (F. Bolle & M. Lehmann-Waffenschmidt, eds.), *Surveys in Experimental Economics*, Contributions to Economics, pp. 95–132, Physica-Verlag HD.

Henning-Schmidt, H., Rockenbach, B. & Sadrieh, A. (2005). *In search of workers' real effort reciprocity – a field and a laboratory experiment*, Discussion Paper 55, GESY.

Jaro, M.A. (1989). Advances in record linkage methodology as applied to the 1985 census of tampa florida. *Journal of the American Statistical Society*, *84*(406), pp. 414–420.

Knutzen, H. (1999). *hkgerman wordlist*, Technical report, Christian-Albrechts-Universität zu Kiel.

Lindeman, R., Merenda, P. & Gold, R. (1980). *Introduction to Bivariate and Multivariate Analysis*, Glenview IL: Scott, Foresman.

Niederle, M. & Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, *122*(3), pp. 1067–1101.

Niederle, M. & Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, *24*(2), pp. 129–44.

Raven, J., Raven, J.C. & Court, J.H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4: The Advanced Progressive Matrices.*, San Antonio, Texas: Harcourt Assessment.

Schooler, J., Ohlsson, S. & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, *122*(2), p. 166.

Spearman, C. (1923). *The nature of intelligence and the principles of cognition*, London: Macmillan.

Spearman, C. (1927). *The Abilities of Man: Their Nature and Measurement*, New York: Macmillan.

Sternberg, R. (2006). The nature of creativity. *Creativity Research Journal*, *18*(1), pp. 87–98.

Styhre, A. (2008). The element of play in innovation work: The case of new drug development. *Creativity and Innovation Management*, *17*(2), pp. 136–146.

van Dijk, F., Sonnemans, J. & van Winden, F. (2001). Incentive systems in a real effort experiment. *European Economic Review*, *45*(2), pp. 187–214.

Winkler, W.E. (1990). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, pp. 354–359.

Ziegelmeyer, A., Schmelz, K. & Ploner, M. (2011). Hidden costs of control: four repetitions and an extension. *Experimental Economics*, pp. 1–18.

A Lettersets

A.1 A British 75%-quantile letterset

This letterset is similar to the German lettersets that we used in the experiment. The only difference is that it has been built with the British ispell dictionary.

We generated 100 000 random lettersets and calculated for each letterset the number of achievable points (here 7049), the number of words (here 528) and the similarity index[19] (here 0.888156). We restricted our attention to lettersets which were close (within 1% margin) to the 75% quantile for points. This is why we call this letterset a "75%-quantile letterset". Similarly we restrict ourselves to lettersets which are within 1% quantile margin for words and similarity of words. Hence, if there are any systematic differences among our lettersets these differences will be small.

letters	points	words	similarity within
`accdeeeginst`	7049	528	0.888156

a ac acts aden aeneid ag agnes agni andes angie as at ats c ca cage cain cains candice case cd ci cid cs d dan dane danes dante dean dec decca deccan dee deena degas dena deng denis denise di diane dina dis e east ed eden edens edna eng enid es etna g ge gte ga gaines gates gd ge gen gena gene genet gide gina i ian ida in ina inc inca incas ind ines inge it n na nat nate nd ne ned ni nice nita s sade sadie san sand sang sat sc se sean sec sega seine sen senate sendai seneca set sgt si sian sid sn snead st staci stacie stan stein stine t ta tad taine tc ted ti tia tide tina ting a accede accedes acceding accent accented accents accident accidents ace aced aces acetic acid acids acing acne act acted acting acts ad ads aegis age aged agencies agent agents ages aid aide aides aids an and ands angst ani anise aniseed ant ante anted anteed antes anti antic antics antis ants as ascend ascent ascetic aside at ate ates c cacti cad cadence cadences cadet cadets cadge cadges cads cage caged cages cagiest can candies cane caned canes cans cant canted cants case cased casein casing cast caste casted casting cat cats cease ceased ceasing cede cedes ceding cent cents cite cited cites cs d dais dance dances date dates dating dean deans decant decants decease deceasing deceit deceits decencies decent deice deices deign deigns den denies dens dense dent dents descant descent desiccate design designate destine detain detains dice dices dicta die dies diet diets dig digest digs din dine dines ding dings dins dint dis disc distance e ease eased easing east eat eaten eating eats edge edges edgiest edict edicts edit edits enact enacted enacts encase encased end ends entice enticed entices es eta g gad gads gain gained gains gait gaits gas gate gated gates gee geed gees geese gene genes genetic genetics genie genies gent gents get gets giant giants gin gins gist gnat gnats gs i ice iced ices id idea ideas ides ids in incest ingest ingested ins insect inset instead is it its n nag nags neat need neediest needs negate negated negates negs nest nested net nets nice nicest niece nieces nit nits nee s sac sad sag sage said saint sand sane saned sang sat sate sated sateen satin satined sating scad scan scant scanted scat scene scened scenic scent scented science sea seat seated seating secede seceding sect sedan sedate sedating sedge see seed seeding seeing seen senate send sent set sic side siege sign signed signet sin since sine sing singe singed sit site sited snag snide snit stag stage staged staid stain stained stance stand stead steed stein steined sting seance t taces tad tads tag tags tan tang tangies tangs tans tea teaed teaing teas tease teased teasing tee teed teeing teen teenage teenaged teens tees ten tend tends tens tense tensed ti tic ticced tics tide tides tie tied ties tin tine tined tines ting tinge tinged tinges tings tins ts

[19] We used the fstrcmp form GNU Gettext 0.17 to calculate for each word the similarity to the most similar word in the set.

A.2 A German 75%-quantile letterset

This is one of the lettersets we used in the experiment. We generated 100 000 random lettersets and calculated for each letterset the number of achievable points (here 5585), the number of words (here 330) and the similarity index (here 0.888436). We restricted our attention to lettersets which were close (within 1% margin) to the 75% quantile for points. This is why we call this letterset a "75%-quantile letterset". Similarly we restrict ourselves to lettersets which are within 1% quantile margin for words and similarity of words. Hence, if there are any systematic differences among our lettersets these differences will be small.

letters	points	words	similarity within
`accehhikllst`	5585	330	0.888436

ach achilles achse achsel acht achte achteck achtecks achtel achtes achtle ahle ai akt akte aktie akts alice alices all all alle alles alls als alt alte altes asche asket ast at ca cache caches call calls cellist ch chalet chalets chate chi chic chice chices chicste chile cia echt eh eilst eilt eis eiskalt eklat elch elchs eli elias elis es esc et etc eth ethik ethisch hacke hackst hackt hackte hai haie haies hais hake hakst hakt hakte hall halle halls hallst hallt hallte hals halt halte hasche hascht haschte hase haskell hast haste hat he hecht hechts heck hecklicht hecklichts hecks heckst heckt hehl hehlst hehlt heil heilst heilt hektisch hell hellst hellt hielt hit ich ist it kachel kahl kahle kahles kahlheit kai kais kali kalis kalt kalte kaltes kastell keil keils keilst keilt kelch kelchs kiel kiels kies kille killst killt killte kiste kit kits kitsch klatsch klatsche kleist kt lach lache lachs lachse lachst lacht lachte lack lacke lackes lacks laiche laichst laicht laichte laie las lasche last laste latsche least lech lechs leck lecks leckst leckt leica leicht leihst leiht leis lest licht lichte lichts lieh liehst lieht lies liest lila lisa list liste lsi lt sache sachlich sachliche sacht sachte sack sacke sackt sackte sah saht saite schach schacht schachtel schah schal schale schalheit schalk schalke schalkheit schall schalle schallt schallte schalt schalte scheck scheich scheit schellt schi schicht schichte schicke schickt schickte schielt schilt schlacht schlachte schlacke schlackt schlackte schlecht schleckt schleicht schlich schlicht schlichte schlick seht sei seicht seil seilt seit sek sekt set sh shell sich sichel sicht sichte sie siech siecht sieh sieht siel skat sketch ski st stach stachel stachle stack stahl stak stall stck steak steil stich stiche stichel stichle sticke stiel stil stile still stille taille takel takels takle tal tales talk talks tals tasche task teich teichs teil teils tel tick ticke ticks tisch tische

B Composition of participants

Figure 10 Composition of participants