How do Incentives affect Creativity?*

Katharina Eckartz† Oliver Kirchkamp‡

January 22, 2020 - revision 176

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 three others tasks which do not require creative thinking. In the first experimental series we find in all tasks that incentives seem to have very small effects and that differences in performance are predominantly related to individual skills. Subjects exerted suprisingly high efforts irrespectively of how they were compensated. In a second experimental series, we focus on two potential explanations: first, subjects might exert effort simply because they enjoy working on the tasks. Second, subjects might exert effort because they feel obliged to do so or because they do not have opportunity costs of working. These questions are crucial to better understand the robustness of experimental results and also to be eventually able to transfer the results to the world outside the laboratory. We replicate our earlier results: in the baseline treatment we do not find effects of incentive schemes on the output. Decreasing the attractiveness of the tasks, we also do not observe differences between the incentive schemes. When we introduce, however, a paid outside option, the efforts are higher in the performance-dependent pay treatments than under flat payment. The size of the effect differs between the tasks, the direction is, however, the same.

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

JEL Classification: C91, J33

*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

1. Introduction

Innovation and creativity are receiving increasing attention in research and business. They are essential for the success of companies in the competitive economy.[1] Their importance has also been recognised by governments who are concerned with the success of their entire economy.[2] Following Huhtala & Parzefall (2007, p. 300), "innovativeness requires creativity". In a similar vein, Amabile (1996, Chapter 8) defines innovation as the "successful implementation of creative ideas by an organisation."[3]

In this paper we ask how employee-creativity is affected by incentives.[4] Incentives are a potentially influential factor under the control of the firm. While it is hard to examine this mechanism with field-data, the incentive-research in experimental and behavioural economics has mainly focused on stated effort experiments.[5] Real effort tasks are used in the laboratory mainly in the context of production tasks which are cognitively undemanding and do not require creativity.[6] We attempt to close this gap and examine the impact of different payment schemes on performance in a creative, real effort task.

Standard microeconomic labour supply theory suggests that people will provide more effort under performance pay. If a cognitive or creative activity is considered as costly (see Camerer & Hogarth, 1999, for a discussion) performance pay would also stimulate creative efforts. However, in the field sometimes incentives seem to 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, an observation which is referred to as "chocking under pressure".

In laboratory experiments with simple real effort production tasks which usually find a positive impact of incentives on output. 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.

Financial incentives in the lab, however, do not always increase performance. Gneezy & Rustichini (2000) find that payments for performance in an IQ-test actually decrease perfor-

[1] In research the importance of innovation is demonstrated, for example, by the specialised journal *Creativity and Innovation Management* which was founded in 1992. In business the importance is stressed, amongst others, in a survey by McKinsey: 70 percent of the interviewed leaders saw innovation among their top three priorities of driving growth (Barsh, Capozzi & Mendonca, 2007).

[2] To monitor innovative activities, governments set up surveys and committees. The European Union set up the "Community Innovation Survey" in 1992 (www.europa.eu). In the US the secretary of commerce established an "advisory committee on measuring innovation" (www.esa.doc.gov).

[3] Similarly, West (2002) distinguishes between idea generation (creativity) and implementation (innovation).

[4] In business a discussion emerged on how to set the conditions to achieve an optimal level of employee creativity, see e.g. DiLiello & Houghton (2008). Their focus is on the discrepancy between creative potential and practiced creativity.

[5] In a stated effort experiment subjects select an "effort level" from a table which is associated with pre-specified costs. Subjects do not actually exert effort. This type of task has been used in many gift exchange experiments; for an overview see Gächter & Fehr (2002).

[6] Lezzi *et al.* (2015) list a number of real effort tasks and give references in their introduction.

mance 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. Ariely *et al.* (2009) perform a controlled experiment in India where they find that performance can decrease when incentives are high.

A number of survey articles summarise the results of earlier experiments examining the effects of different payment schemes and try to identify general pattern: Camerer & Hogarth (1999) review a large number of experimental studies on the effects of performance-based incentives. They find no effect on mean performance. Camerer & Hogarth observe, however, that the effects differ between the analysed tasks. Bonner *et al.* (2000) focus in particular on how different incentives work in different task types. They mainly focus on task complexity and conclude that positive incentive-effects were only found in half of the studies. In particular, positive effects were mainly observed in simple tasks. In an attempt to compare the experimental practices between economics and psychology, Hertwig & Ortmann (2001) discuss, among others, the effects of financial incentives. They examine a number of different tasks. When they observe a difference, in the majority of cases incentives lead to a higher performance. Last, Prendergast (1999) looks at field studies and finds positive effects of pay-for-performance in tasks in which performance was easily measurable.

What should we expect for an experiment on creativity?[7] As long as participants are not intrinsically motivated, participants should perform better under performance pay as compared to a flat payment. With a creative task, however, intrinsic motivation might be relevant. Introducing incentives could then crowd out intrinsic motivation and could be counterproductive.[8] In fact, Fessler (2003) finds that pay-for-performance decreases the perceived task attractiveness.

Even if motivation is not crowded out by financial incentives, higher effort is not necessarily linked to higher output. This holds in particular for tasks in which subjects have to think "unorthodoxly" (Camerer & Hogarth, 1999).[9] It is quite likely that creative tasks will often fall into this category, in particular when we consider the importance that some

[7] There is a large body of psychological research on creativity amongst others by Amabile and her co-authors as well as by Sternberg and co-authors. Teresa Amabile's conceptual definition of creativity is: "A product or response will be judged as creative to the extent that (a) it is both novel and appropriate, useful, correct or valuable response to the task at hand, and (b) the task is heuristic rather than algorithmic" (Amabile, 1996, p. 35). Robert Eisenberger, following a different approach, puts large focus on divergent thinking (e.g. Eisenberger & Selbst, 1994) which was also stressed in the early research on creativity which followed psychometric approaches (Guilford, 1950 and Torrance, 1968).

The psychological research on creativity 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 amongst 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. Amabile (1996) and Sternberg (1999) provide overviews about the different branches of psychological creativity research.

[8] Motivational crowding out goes back to Deci (1971). Similarly, imposing other controls, like minimal effort levels and monitoring (amongst others Falk & Kosfeld, 2006 and Ziegelmeyer *et al.*, 2011) have been identified as potentially reducing effort.

[9] Ariely *et al.* (2009) argue that a too high motivation can increase arousal too much and thereby hamper performance. This effect is known as the "Yerkes-Dodson law" (Yerkes & Dodson, 1908).

psychological theories put on “divergent thinking” (see, e.g., Eisenberger & Selbst, 1994).[10]

Besides the inherent reward of working on the experimental task, there are a number of other potential reasons that might lead the experimental subjects to exerting considerable effort. The first possible explanation is that subjects have a high general or task-specific motivation to perform well. Either because they enjoy working on the tasks and the pleasure of the task offsets their cost of working on it, because they feel challenged and enjoy taking up this challenge, or because of experimental effects. It might be, for example, that the experiment is, in particular in experiments which mainly have student subjects, perceived as an exam situation. The second potential reason is that subjects feel a moral obligation, for example to the experimenter, to exert effort knowing that they will receive a compensation for participating in the experiment. Further explanations are that subjects work on the tasks because they do not have signifficant opportunity costs of working and they also do not have outside options. Consequently, subjects might work only to prevent themselves from being bored. A last explanation, specific to within-subjects design is that subjects use the non-contingent pay periods to practice for potential later contingent pay periods.

In the second experiment we will investigate two of these possible explanations. We will first reduce work-motivation by making the tasks less attractive. *Potential ways of achieving this goal are making the tasks more difficult, frustrating or boring. The first two characteristics are linked: making tasks moderately difficult might constitute a challenge, while making them very difficult might make working on them very frustrating.* Second, we introduce a paid outside option and thereby aim to investigate jointly the introduction of opportunity cost and the legitimacy of not working. The implemented outside option does not aim at reducing subjects’ possibly existing boredom. With respect to the task characteristics,Camerer & Hogarth (1999) discuss the possible relation between task difficulty and the potential of incentives to influence performance. Camerer & Hogarth argue that this potential is highest for tasks with intermediate difficulty. When the task is very easy, high performance is achievable easily ("floor effect"). When the task is very difficult, even though subjects might respond to incentives and increase efforts, it is likely that this increased effort is not resulting in improved results ("ceiling effect"). The importance of skills, and in particular the discrepancy between the needed and the possessed skills, is stressed by Bonner *et al.* (2000). The authors find positive effects of incentives mainly in simple tasks. Furthermore, next to the mere effect of the matching of skills,Brase (2009) argues that also the subjects’ perceived likelihood of increasing output with increased efforts will contribute to the observability of incentive-effects: when subjects rate this likelihood too low, they might give up too early. Ariely *et al.* (2009) analyse the effects of financial incentives in a number of diferent task types and with different stake-sizes. They find that the effects are, in fact, dependent on the tasks and task characteristics. Relevant for this experiment is especially their finding that incentives have a positive effect in a key-pressing task but negative effects in a number adding task. Task attractiveness also links to the research by Deci and colleagues on the effects of financial rewards on intrinsic motivation. When the task attractiveness is too high, no or even detrimental effects of financial incentives are to be expected (e.g. Deci *et al.*, 1999). Besides,

[10]Divergent thinking is understood as the “production of varied responses” in a task that has different alternative solutions (Eisenberger & Selbst, 1994, p.1116).

Bailey & Fessler (2011) looked at both task attractiveness and difficulty in one study. The authors find that pay-for-performance, in their case a piece-rate, was only effective in the unattractive and the non-complex tasks. When manipulating the task enjoyability, we tried to decrease task enjoyability in the creative task by making the task more difficult, however still solvable. We modified the effort task such that it is easier, but not very attractive. With respect to the introduction of outside options, different techniques have been applied in the experimental economic literature so far. These techniques varied on the degrees of attractiveness and control: Dickinson (1999) offered subjects the possibility to go home after they completed a minimum number of tasks. He refers to this possibility as "off the job leisure". He extends the classical labor-leisure-model and considers both the number of hours worked, as well as the effort during the time worked ("on and off the job leisure"). An alternative approach is taken by the studies that introduce time-out buttons in the experiment: Mohnen *et al.* (2008) included an incentivised time-out button in their study. Their focus was on peer pressure in team-work resulting from inequality aversion. In their study students were working independently on a real effort task while their earnings were shared in a 2-person-team. They introduced the time-out button to include an opportunity cost of working. Pushing the time-out button was interpreted as working for the private account. Similarly, Blumkin *et al.* (2010) used an incentivised stop-button. Their set-up worked completely with consumption goods (food-vouchers for produced units and drink-vouchers for leisure-time). The focus of their study was to test whether a labor income tax and an equivalent consumption tax lead to an identical labor-leisure allocation. Furthermore, in the study of Gamage *et al.* (2010) subjects had to make a decision to work for the next 9 minutes and be compensated accordingly or to receive a fixed fee and watch preselected videos instead. The authors' focus was on the effect of different descriptions of aftertax income on the willingness to work and the amount worked by those who chose to work. A study which aimed at providing attractive outside options to the subjects was conducted by Corgnet *et al.* (2013). The authors developed a platform on which subjects could easily switch between real effort and real leisure ("on the job leisure"), implemented as surfing on the internet. The authors' focus was on comparing individual and team incentives as well as on the effect of monitoring. While individual incentives originally outperformed team incentives, the authors found a large positive effect of peer monitoring on efforts. From this overview, we see that a number of different outside options have been applied in the literature so far.

Our focus is on the difference in subjects' performance under different incentive schemes, once in a pure effort (non-creative) task and once in a creative task. As will be seen, regardless whether subjects have in incentive or not, on average subjects exert a substantial effort.

Why do subjects exert so much effort? We modify the experimental set-up accordingly: on the one hand we introduce a paid outside option. On the other hand we decrease the task-attractiveness. The design in the second experimental series will be closest to the one of Mohnen *et al.* (2008). This design allows subjects to push a button in order to switch between pause and working. It is, hence, easy to observe how much time subjects spend on pause and to can compensate them accordingly. In addition, the pause button is easy to implement in our experimental setting and allows us to keep a lot of experimental control.

The experimental economic literature which is studying the effects of financial incentives on subjects' performance reached inconclusive results. Here we compare the effects of

three different payment schemes in three different non-creative tasks. We find it surprising that treatment effects are very small. This outcome is not the result of subjects that are not performing: subjects do exert a significant amount of effort in our experimental tasks regardless of the treatment. We find it, in particular, surprising to see good performanace and no reaction to incentives even for number adding task. The emerging question is: why do experimental participants exert substantial effort when they do not have a financial incentive to do so. There are a number of potential reasons that might lead the experimental subjects to exert considerable effort.

The first possible explanation is that subjects have a high general or task-specific motivation to perform well. Either because they enjoy working on the tasks and the pleasure of the task offsets their cost of working on it, because they feel challenged and enjoy taking up this challenge, or because of experimental effects. It might be, for example, that the experiment is perceived as an exam situation.[11]

The second potential reason is that subjects feel a moral obligation, for example to the experimenter, to exert effort knowing that they will receive a compensation for participating in the experiment. Further explanations are that subjects work on the tasks because they do not have significant opportunity costs of working and they also do not have outside options. The only available outside option in the laboratory is to sit and do nothing, maybe letting your thoughts flow. Consequently, subjects might work only to prevent themselves from being bored.

A last explanation, specific to our within-subjects design in our first study, is that subjects use the non-contingent pay periods to practice for potential later contingent pay[12] periods.[13]

We will investigate the first two explanations. First, we aim at changing subjects' work-motivation by making the tasks less fun. Potential ways of achieving this goal are making the tasks more difficult, frustrating or boring. The first two characteristics are linked: making tasks moderately difficult might constitute a challenge, while making them very difficult might make working on them very frustrating. Second, we introduce a paid outside option and thereby aim to investigate jointly the introduction of opportunity cost and the legitimacy of not working. The implemented outside option does not aim at reducing subjects' possibly existing boredom.

With respect to the task characteristics, Camerer & Hogarth (1999) discuss the possible relation between task difficulty and the potential of incentives to influence performance. Camerer & Hogarth argue that this potential is highest for tasks with intermediate difficulty. When the task is very easy, high performance is achievable easily ("floor effect"). When the task is very difficult, even though subjects might respond to incentives and increase efforts, it is likely that this increased effort is not resulting in improved results ("ceiling effect"). The importance of skills, and in particular the discrepancy between the needed and the possessed skills, is stressed by Bonner *et al.* (2000). The authors find positive effects of incentives mainly

[11] After all, most of our subjects are students, and, thus familiar with the situation of an exam.

[12] Contingent pay and performance pay will be used interchangeably to subsume linear and tournament payment mechanisms.

[13] The design in our first study consisted of 7 stages, in which all incentive schemes were conducted within subjects. The last stage was a self-selection stage. Treatment information were only provided directly before every stage.

in simple tasks. Furthermore, next to the mere effect of the matching of skills, Brase (2009) argues that also the subjects' perceived likelihood of increasing output with increased efforts will contribute to the observability of incentive-effects: when subjects rate this likelihood too low, they might give up too early. Ariely *et al.* (2009) analyse the effects of financial incentives in a number of different task types and with different stake-sizes. They find that the effects are, in fact, dependent on the tasks and task characteristics. Relevant for this study is especially their finding that incentives have a positive effect in a key-pressing task but negative effects in a number adding task. Task attractiveness links to the research by Deci and colleagues on the effects of financial rewards on intrinsic motivation. When the task attractiveness is too high, no or even detrimental effects of financial incentives are to be expected (e.g. Deci *et al.*, 1999). Besides, Bailey & Fessler (2011) looked at both task attractiveness and difficulty in one study. The authors find that pay-for-performance, in their case a piece-rate, was only effective in the unattractive and the non-complex tasks.

When manipulating the task enjoyability, we tried to decrease task enjoyability in the creative task by making the task more difficult, however still solvable. Moreover, we modified the non-creative task such that it is easier, but not very attractive.

The rest of the paper is organised as follows: Sections 2 presents the design of the experiment. In Section 3 we present our hypotheses. Section 4 presents results. Sections 5 and 6 provide a discussion and conclusion.

2. Experiment

In our treatments we combine different tasks with different types of incentives. In Section 2.1 we describe the tasks. In Section 2.2 we describe the incentives. Section 2.3 explains how tasks and incentives are combined in treatments. In Section 2.4 we describe the implementation of the experiment.

2.1. Tasks

Table 1 gives an overview of the different tasks in the experiment. To study the different incentive schemes (see Section 2.2) we use creative tasks and non-creative tasks. We also vary the difficulty or attractiveness of the tasks. The different incentive schemes are compared within-subject. Also creative and non-creative tasks are compared within-subject. Task difficulty is compared accross-subject.

2.1.1. Paid pause

In this study we want to find out how incentives work in creative tasks. One consequence of an incentive might be that workers spend more time with the incentivised task. The effect of incentives might also depend on the opportunity cost of work or the available outside option.

To model outside options, different approaches have been used in the experimental economic literature so far. These approaches differ in the attractiveness of the outside option but also in the experimental control: Dickinson (1999) offered subjects the possibility to go

Table 1 Experimental tasks

Treatment	Pause	Creative task	Non-creative task	Participants	Sessions
BL+IQ	-	creating words min. word length: 1	IQ-task	166	10
BL	-	creating words min. word length: 1	adding 5 2-digit numbers	102	6
Pause	paid pause	creating words min. word length: 1	adding 5 2-digit numbers	43	3
DiffBL	-	creating words min. word length: 6	counting *1*s in a 5×5 matrix of *0*s and *1*s	51	3
DiffPause	paid pause	creating words min. word length: 6	counting *1*s in a 5×5 matrix of *0*s and *1*s	49	3

home after they completed a minimum number of tasks. He refers to this possibility as off the job leisure. He extends the classical labor-leisure-model and considers both the number of hours worked, as well as the effort during the time worked ("on and off the job leisure"). An alternative approach is taken by the studies that introduce time-out buttons in the experiment: Mohnen, Pokorny & Sliwka (2008) included an incentivised *time-out button* in their study. Their focus was on peer pressure in team-work resulting from inequality aversion. In their study students were working independently on a real effort task while their earnings were shared in a 2-person-team. They introduced the time-out button to include an opportunity cost of working. Pushing the time-out button was interpreted as working for the private account. Similarly, Blumkin, Ruffle & Ganun (2010) used an incentivised *stop-button*. Their set-up worked completely with consumption goods (food-vouchers for produced units and drink-vouchers for leisure-time). The focus of their study was to test whether a labor income tax and an equivalent consumption tax lead to an identical labor-leisure allocation. In the study of Gamage, Hayashi & Nakamura (2010) subjects had to make a decision to work for the next 9 minutes and be compensated accordingly or to receive a fixed fee and watch preselected videos instead. The authors' focus was on the effect of different descriptions of after-tax income on the willingness to work and the amount worked by those who chose to work. A study which aimed at providing attractive outside options to the subjects was conducted by Corgnet, Hernan-Gonzalez & Rassenti (2013). The authors developed a platform on which subjects could easily switch between real effort and real leisure ("on the job leisure"), implemented as surfing on the internet. The authors' focus was on comparing individual and team incentives as well as on the effect of monitoring. While individual incentives originally outperformed team incentives, the authors found a large positive effect of peer monitoring on efforts.

From this overview, we see that a number of different outside options have been applied in the literature so far. The design in our second study is closest to the one of Mohnen *et al.*

(2008). As in their design we allow subjects to switch between pause and working.

2.1.2. Creative task

It is not obvious to find a task that requires creative thinking and that is, at the same time, suitable for an experiment. To be suitable for an experiment, it must be possible to assess the quality of the solution of the task quickly and easily. Also, the task must remain interesting when it is repeated. Repetition must be associated with only limiteted learning-effects. 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 not be repeated. Once a participant has understood the problem, the solution can, with or without incentives, quickly be applied again.[14]

Open tasks like "painting a creative picture" remain interesting with repetition. However, for the experimenter it is hard to judge the quality of the solutions. In particular, it is hard to apply incentives which are based on quality. Quality of submissions can be assessed with the help of experts (Amabile, 1996), other researchers, a larger group of students, or a web based tool (Girotra *et al.*, 2009). These assessment, however, take too much time in a repeated laboratory experiment.[15] Hence, here we will use tasks that can be quickly and mechanically rated by the computer. Bradler *et al.* (2019) find for a different task (unusual uses) that tournaments seem to have a positive effeect.

Word task: In our study we use a word creation task[16] as our creative thinking task: participants are presented with an alphabetically ordered letterset, consisting of 12 letters, e.g. `accdeeeginst`. Their task is to create 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 2 gives some examples of words that can be constructed with these letters and the resulting points.[17] Appendix A.7.1 shows all English words that a participant could find for the above letterset. Appendix A.7.2 shows all German words for a similar letterset.

Such a "word task" has many aspects of a creative task and mimics creative innovation quite well. 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

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

[15] In tasks like these the quality of the submitted solutions is usually rated by a jury in psychological research. Bradler *et al.* (2013) used the "unusual uses" task of the Torrance Test of Creative Thinking Torrance (1968) and found a way (based on a pre-test) to relatively quickly rate the submitted solutions such that it was possible to make experimental compensation performance dependent.

[16] 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. In creativity research two studies used similar tasks: Eisenberger *et al.* (1999) presented participants with long words and subjects had to create shorter words out of these. Stone (1971) gave his participants a letterset. In Stone's experiments subjects had to create *new* words from the letterset. The created words were evaluated by a jury afterwards.

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

Table 2 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

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. It is likely 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.[18]

After a pilot in which we used all 8 lettersets, we dropped the 2 best- and the 2 worst-scoring ones. Table 3 shows which lettersets were used in the final experiments. Which lettersets were used was depending on the treatment. 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 correctly entered words were shown as a list on the screen. Participants were not informed about how many points they had accumulated. In the *treatment modification* of the second experimental series a minimum word length of 6 letters was introduced to increase the difficulty of the task. This minimum word

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

length was based on the results of a pilot experiment.[19]

Difficult word task Incentives might be more necessary or more influentual when the task is more difficult or less attractive. Therefore, we consider most of our tasks in a more or less attractive version.

In the enjoyable (BL) version of the treatment all words would count. In the less attractive version of the task (DiffBL) only words with a minimum length of 6 letters would count.

Table 4 in the Appendix displays the results of the manipulation check: subjects find the creative task with a minimum number of letters indeed less enjoyable and more difficult than the BL task. Task importance is not influenced.

2.1.3. Alternative tasks

For the non-creative task we compare an IQ-task with a number-adding task and a number-counting task.

IQ task As a more attractive[20] non-creative task we use an IQ-task which is 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 10) in the Appendix.

Number adding task As a less attractive non-creative task we use a number adding task, similar to the one used by Niederle & Vesterlund (2007): for five minutes participants had to add five two-digit numbers.[21] They were allowed to use scratch-paper for their calculations. Moreover, after 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 in the pure IQ-task.

[19] We aimed to decrease the task enjoyability; at the same time we wanted to keep the task difficulty still intermediate (as discussed in Camerer & Hogarth (1999). Based on the pilot experiment, a minimum number of 7 or 8 letters seemed to be too difficult for the subjects, while with a minimum of 6 letters subjects still managed to create a substantial number of solutions.

[20] A manipulation check is provided in Appendix A.3.

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

Counting task As an even less attractive non-creative task we use a counting task.

In the counting task subjects see a 5×5 matrix consisting of 0s and 1s. Their task is to count the number of *1*s in that matrix (similar to the task used in Houser, Schunk, Winter & Xiao, 2010).[22] This task was chosen such that everybody can do it and thereby give as little feedback about potentially meaningful skills as possible. Moreover, it is hard to imagine that working on this task is particularly rewarding or fun.[23]

Table 5 in the Appendix displays the results of the manipulation check: Indeed, subjects find counting ones less challenging than adding numbers. The non-creative tasks are also perceived as less enjoyable than the creative task.

2.1.4. Questionnaire:

At the end of the experiment participants answered a questionnaire including questions on participants' interest in the two different 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).[24] The distribution of risk preferences is shown in Figure 17 in Appendix A.8.

Last, as a manipulation check, in the second experimental series subjects were asked how much they enjoyed working on the tasks, how difficult they rate the experimental tasks, and how important it was for them to perform well on the tasks ("task importance"). In the treatments that contained an outside option, the questionnaire also included an open-ended question in which subjects were asked to describe how they used the Pause option to possibly get some insights into the motivations underlying subjects' behaviour.

2.2. Incentive schemes

We are interested in participants' performance under different payment schemes in a given time. In these experimental series we compared three incentive schemes: a *flat fee* regime, a *linear* payment regime and a *tournament*.[25] All parameters were calibrated such that the expected payment for the experiment, which lasted for approximately one hour, was about 10€. This was considerably more than the average hourly wage of a student assistant at the University of Jena at that time. In contrast to other studies who focus on the provided

[22] Ariely *et al.* (2009) found a positive effect in a key-pressing task, which might be comparatively interesting to the *counting 1s task*. They find negative incentive effects in an adding numbers task.

[23] In fact, Houser *et al.* (2010, p. 5) designed their task with the goal "to be boring".

[24] Dohmen *et al.* (2011) included the question in the 2004 wave of the German Socio Economic Panel. This measure consists of a 11-point scale, reaching from 0 (being very risk-averse) to 10 (being very risk-loving). They found this question to be correlated with real risk-taking behaviour while a lottery choice did not predict real risk-taking behaviour as well as this simple question.

[25] Tournaments are discussed extensively in the literature (e.g. by Bonner *et al.*, 2000, van Dijk *et al.*, 2001, Harbring & Irlenbusch, 2008 and Prendergast, 1999). Its practical advantage is that they are easily implementable as one needs only information about the relative performance. Moreover, tournaments circumvent a problem that might arise also outside the laboratory under different incentive schemes namely underreporting of true performance. Potential disadvantages of tournaments are that some people might give up and that it might hinder cooperation in teams (all discussed in Prendergast, 1999).

working-time we focus on the effects on subjects' performance in a given time. For higher effort to result in higher output, the match between the task difficulty and the subjects' skill has to be good enough (Camerer & Hogarth, 1999). We believe that our subject pool consisting predominantly of students satisfies this criterion.

During the experiment participants received points for correct solutions. At the end of the experiment one of the experimental stages was randomly selected for payment to prevent participants from hedging between stages. The respective number of points was converted into Euros with an exchange rate of 1 point = 0.04€. In the *flat* scheme participants received 250 points (=10€) irrespective of their performance. The payment in the *linear* incentive scheme was dependent on the task: In the creative task, the instructions asked the participants to create as many and as long words as possible. In the two performance pay conditions, we rewarded the increasing difficulty to construct long words with more than proportionally more points. More specifically, participants received for every correctly created word 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 Table 2). In the other tasks 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 with 25 points and in the counting task every correctly solved problem was awarded with 6 points.[26] In the tournament the number of acquired points was compared to the points of three other participants for the respective task who faced the same treatment order.[27] A winning participant was awarded 25€ (if that condition was chosen for payment) and a losing participant was compensated with 5€.[28] 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.[29]

2.3. Treatments

The experiment consisted of eight stages[30], each lasting five minutes. In each incentive scheme, participants always started with the creativity task and afterwards solved the non-creative task with the same incentive scheme. We varied the sequence of incentive schemes to rule out order effects. No feedback was given during the experiment. Table 3 provides an

[26] The piece-rate in the IQ-task, the creativity task and the counting 1s 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).

[27] Thus, the number of subjects per session did not have to be a multiple of 4.

[28] Ties were broken randomly by the computer.

[29] 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.

[30] For the sessions where the non-creative task was the IQ task, as well as for 3 sessions where this task was the number adding task we dropped the eigth's stage, i.e. the stage where participants could choose an incentive scheme for the non-creative task

Table 3 Stimuli

	Letterset / non-creative task		
Stage	BL & Pause	DiffBL & DiffPause	Incentive
1	`aceehhinrssä`	`aabeefghllnn`	incentive scheme 1
2	IQ or numbers	counting *1*s	incentive scheme 1
3	`aeeeggllmnru`	`ceefiiknnstt`	incentive scheme 2
4	IQ or numbers	counting *1*s	incentive scheme 2
5	`deehhimnnprt`	`aeeehknnsstt`	incentive scheme 3
6	IQ or numbers	counting *1*s	incentive scheme 3
7	`deegilmnnpuw`	`aeeggiilnnns`	self-selection
8	numbers	counting *1*s	self-selection
	questionnaire	questionnaire	

overview.

The last stage of the experiment was a self-selection stage. Participants could choose which of the previously experienced incentive schemes they preferred for the subsequent word creation 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 (see, e.g., Niederle & Vesterlund (2010)). We included the self-selection stage as this allows us to investigate several questions: who selects which incentive scheme, do we find differences in performance following self-selection and, if so, whether this represents sorting. A number of studies analyse determinants of self-selection. Niederle & Vesterlund (2007) find gender differences in the choice of the preferred payment scheme in their number adding task: having to choose between a tournament and a linear payment scheme, 73% of the men and less than half as many women (35%) chose the tournament. Furthermore, Eriksson *et al.* (2009) look in a stated-effort experiment, amongst others, on the impact of risk preferences. The authors find that risk-averse subjects are less likely to enter tournaments.

2.4. Participants and procedures

Participants We conducted 25 sessions with 411 participants. Participants were recruited online with the help of ORSEE (Greiner, 2004). For each session we invited an equal number of men and women so that group-composition effects related to gender are kept as small as possible.[31] In some sessions we still have small deviations from a perfect balance of men

[31] Gneezy *et al.* (2003) found that for women's performance in tournaments the gender-composition of the reference group is of relevance. By inviting an equal number of men and women we, therefore, wanted to keep this potential impact on performance constant across sessions.

and women since not all invited participants showed up. Overall, we had 202 male and 203 female participants. For 6 participants we do not know their gender.[32]

Most (91%) of our 411 participants were undergraduate students from a broad range of fields of studies. The average age of all participants was 23.7 years. The average payment was 10.52€.

In Appendix A.8 we give more information about the characteristics of our subject pool. There, we also give an overview about the responses to the post-experimental questionnaire items. We do not use data from the pilot session.[33]

Procedure Before the experiment started, participants were waiting in the corridor, hence, they were aware of the composition of the experimental group. Yet, nobody in the experiment was aware of the identity or gender of their matching group members.

The experiment was conducted in German. Participants were informed in the invitation that to participate in the experiment they had to speak German as fluently as a native speaker. They also knew that they had to pass a short German language-test prior 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. In addition, participants 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 a native speaker. The average self-reported knowledge of German was 4.9 on a scale from 1 to 5. We also collected information about the knowledge of other languages. The distribution of the language competence for German and the other languages is displayed in Figure 10 in Appendix A.1.

The experiment was programmed browser-based using PHP with a MySQL database and an Apache server. All entered words were spell-checked using the German isoword-list (Knutzen, 1999).[34] 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 using the Firefox-Plugin *R-Kiosk*. Participants could not leave the full screen mode. Participants also could not move backwards or forwards in the experiment.

At the end of the experiment one of the eight tasks was randomly chosen to determine the payment. We then distributed receipts. Thereafter, participants exchanged signed receipts against an envelope with their payment. Sessions lasted for about one hour.

[32] The distribution over different sessions is shown in the left graph in Figure 16 in the Appendix.

[33] The pilot session served to calibrate experimental parameters (in particular how many points each correctly solved task was awarded with) so that the expected payment in all tasks was equal to 10 Euro. Moreover, the pilot session was used to calibrate the minimum word length for the creative task in the Diff-treatments. Furthermore, we encountered technical problems during one session. We do not use data from that session. Also, data for one individual participant who encountered a technical problem on her computer is not used.

[34] Based on pretest-results this word-list was extended to include more valid words using the German online dictionary Duden.de.

3. Hypotheses

Classic economic labour supply predicts a positive relation between incentives and performance. We hypothesize the following:

Hypothesis 1 *For all tasks productivity will be higher with incentives than without.*

If a task is per-se intrinsically motivating, then participants would provide a maximal effort already without any external incentive. Making a task less attractive (as in DiffBL or DiffPause) should reduce intrinsic motivation.

Hypothesis 2 (task attractiveness) *With the less attractive tasks the positive effect of incentives is stronger.*

Similarly, a more attractive outside option rivals with intrinsic motivation.

Hypothesis 3 (outside option) *With the presence of a paid pause option the positive effect of incentives is stronger.*

In the last rounds of the experiment participants can choose their own preferred incentive scheme. Here we expect that participants who are better at a task, either because they are more motivated or because they are more able, prefer a steeper incentive.

We should hence, find a higher performance with higher incentives already as a result of the selection of participants into incentives.

Hypothesis 4 *Under self-selection performance will be higher under contingent pay than under flat pay, regardless of the treatment.*

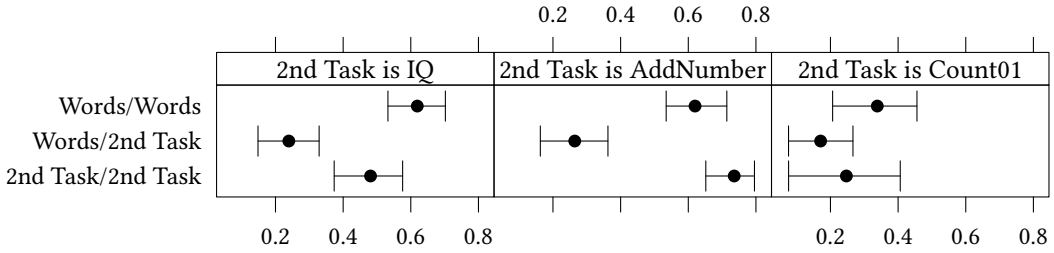
4. Results

4.1. Correlation of task performance

To assess whether we rely on different or rather similar skills with the experimental tasks we show 95% confidence intervals (based on an ABC bootstrap)[35] for correlations of the performance for the different tasks in Figure 1.

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 non-creative tasks is correlated. However, correlation of performance in the word task with performance in the non-creative task is much lower. Though still positive, we can say that creative and both non-creative tasks seem to depend on quite different skills.

Figure 1 Correlation of performance among the different tasks



The segments show 95%-confidence intervals (based on ABC bootstraps).

Figure 2 Performance in the word task

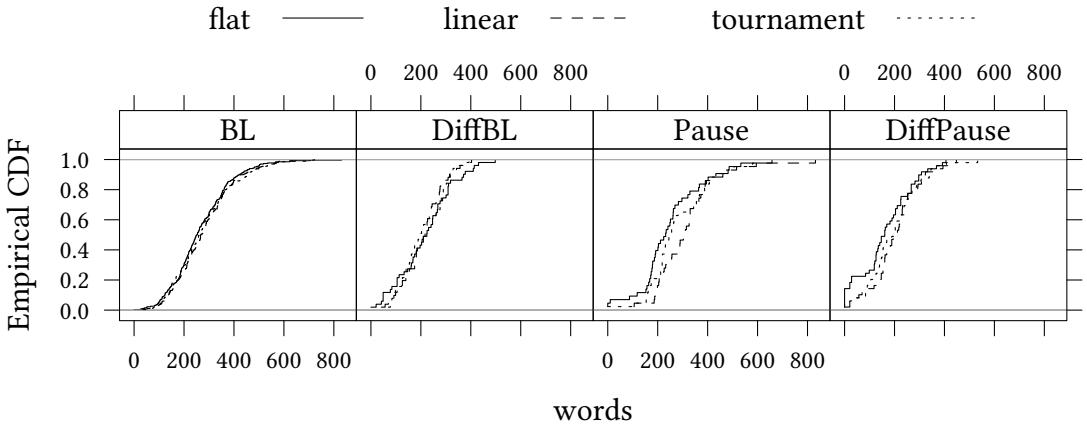
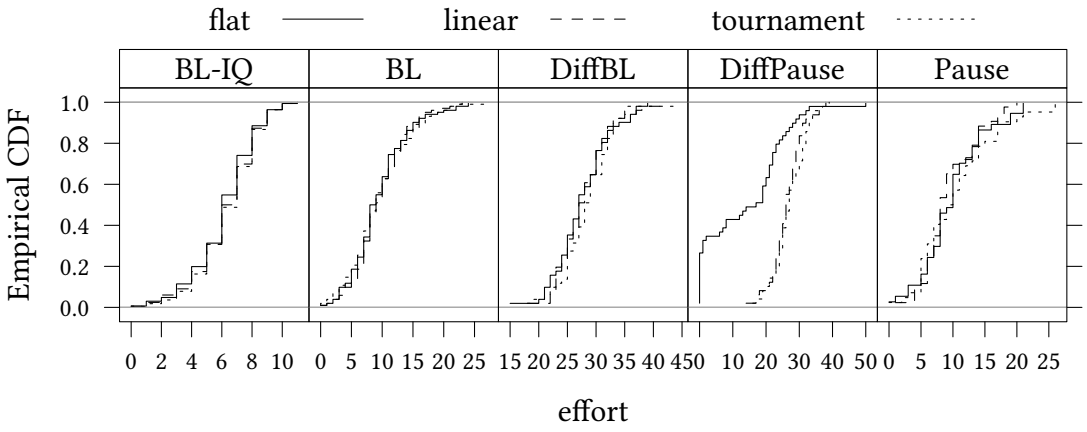


Figure 3 Performance in the effort task



4.2. Performance

In this section we investigate whether incentives have a substantial influence on performance. Performance is measured as the number of acquired points, either in the creative task or in the non-creative task. Figure 2 shows the distribution of performance for the word task. Figure 3 shows the distribution of performance for the effort task. To make the different tasks comparable, performance in the tasks is based on the percentile rank within each task.[36]

We use mixed effects regressions to allow for subject-specific heterogeneity. $\epsilon_{\text{subj.}}$ is a random effect for each participant. ϵ_{stage} capture potential differences between stages. $\epsilon_{\text{subj.},t}$, is the residual. Confidence intervals of the standard deviation of the random effects are displayed in Figure 11 in Appendix A.2. As the distributions of the random effects differ between the treatments and as the separate regressions are more intuitive to interpret we estimate one model per treatment. We estimate the following mixed effects equation:

$$Y = \beta_0 + \sum_{\text{Incent.}} \beta_{\text{inc.}} \cdot d_{\text{inc.}} + \epsilon_{\text{stage}} + \epsilon_{\text{subj.}} + \epsilon_{\text{subj.},t} \tag{1}$$

In this equation the incentive scheme *flat* is the baseline. In the current section Y is performance. Confidence intervals for $\beta_{\text{inc.}}$ are shown in Figure 4.[37] For the baseline treatment the effects are small for the word task but also for the non-creative tasks, they are not significant for both non-creative tasks and only significant in the creative task.

The results are the basis for the following analysis. Hypothesis-tests refer to two-sided alternatives.[38] Hypotheses 2 and 3 assume a direction of effect, thus testing against the one-sided alternative would be justified. Nonetheless, we will stick to using the displayed two-sided p-values as these are more conservative.

Let us look at the first regression in Figure 4. In relation to the size of the intercept[39] ($\beta_0 = 0.483$) the estimated coeffiencts are small in magnitude ($\beta_{\text{lin.}} = 0.0329$), $\beta_{\text{tourn.}} = 0.0195$) and not significantly different from zero. Thus, the result of the previous study can be replicated for the creative task.

Regarding experimental efforts as costs (as discussed by Camerer & Hogarth, 1999) would imply low efforts under flat payments. It is, however, possible that subjects enjoy working on the experimental tasks and, therefore, the non-monetary benefits arising from working on the task itself outweigh the costs of effort which could lead to the observations that we made in the previous study. If this reasoning is relevant, we expect to observe effects of the incentive schemes once the attractiveness of the task is reduced. Therefore, in the experimental manipulation in dimension 1 we try to decrease the task enjoyability by making the creative task more difficult.

[35] Throughout this paper the statistical analysis is done with the statistical software R version 3.6.2 (2019-12-12) (R Development Core Team, 2019).

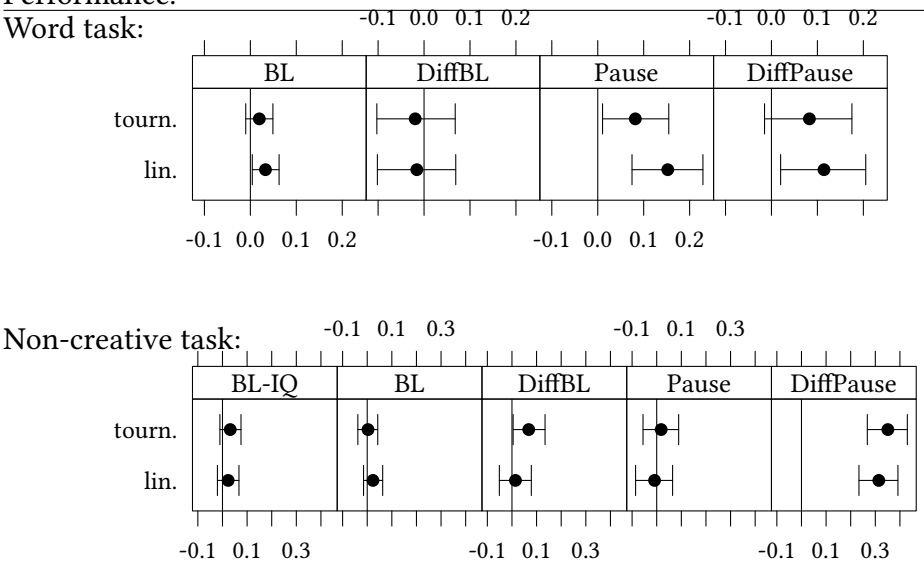
[36] Estimation results for performance as the raw number of points is shown in Figure 13 in the Appendix.

[37] Confidence intervals for fixed effects are based on a normal bootstrap with with 500 replications using `bootMer` from `lme4` 1.1-21.

[38] P-values and confidence intervals were bootstrapped using `bootMer` from lme4 1.1-21 (Bates *et al.*, 2015) with 500 replications.

[39] Here the intercept represents average performance in the BL treatment under flat incentives.

Figure 4 95%-confidence intervals for fixed effects for incentives from Equation (1). Y is Performance.



To make the different tasks comparable, performance in the tasks is based on the percentile rank within each task. Estimation results for performance as the raw number of points is shown in Figure 13 in the Appendix. Random effects are shown in Figure 11 in the Appendix.

It seems that increasing the task difficulty leads to a general drop in performance. For the treatment DiffBL the intercept is lower than in *BL* ($\beta_0 = 0.515$) and second, both performance pay-coefficients are small in magnitude and not significantly different from zero. Hence, we do not find evidence in favour of the reasoning above.

Introducing a paid pause option adresses two points jointly: it should remove a potentially existing feeling of a moral obligation to work and it introduces opportunity costs of working into the experimental set-up. The treatment Pause shows the predicted positive impact of performance pay. Both $\beta_{\text{inc.}}$ have a relatively large magnitude ($\beta_{\text{lin.}} = 0.152$), $\beta_{\text{tourn.}} = 0.0818$) but only $\beta_{\text{lin.}}$ is significant. Comparing the performance in the different incentive schemes with those in BL, it becomes obvious that the observable effect of the incentive schemes is caused by a drop in performance under flat payments. Thus, we find evidence in favour of our reasonig above, however the effect is significant only for linear incentives. Interestingly, performance under flat payments is still significantly larger than zero.

In the Pause treatments also the time that subjects spent on pause is observable. The time on pause was, as expected, different from the theoretical benchmark: the average time on pause under flat payments was 16 seconds, this is about one fifth of the total time they could spend on pause. Moreover, in all incentive schemes the proportion of subjects who do not spend any time on pause is not negligible (Table 8 in the Appendix). This fraction is lowest under flat pay. [40] Introducing a pause option has no impact on work productivity

[40] In Appendix B Table 6 shows that the time on pause was significantly lower under performance pay than under flat pay.

(Figure 5). Thus, the observed performance difference results from the longer time on pause that subjects take under flat payments which results in a drop of performance.

The fourth regression in Figure 4 tackles treatment differences when the task is less enjoyable and at the same time a pause option is introduced. Also here, a positive effect of performance pay is observable. Like in the Pause-treatment, the difference is caused by the drop in performance under flat payments. As the task characteristics are different from the Pause treatment, the sizes of $\beta_{\text{tourn.}}$ and $\beta_{\text{lin.}}$ in regressions 3 and 4 are not directly comparable, here regression 2 would be the relevant reference for regression 4. The work productivity is not significantly different in the performance pay treatments compared to the flat payment (The fourth regression in Table 9 in the Appendix) .[41]

4.3. Non-creative task

The lower part of Figure 4 shows estimation results for Equation 1 for the non-creative task. We compare three different tasks which differ mainly in their cognitive demand: an IQ-task, adding numbers and counting numbers. For the two less demanding tasks, adding number and counting numbers, we also add a pause option.

The second regression shows the estimation results when the task is the *counting 1s* task. The regression shows that the two performance pay coefficients are not significantly different from zero. Thus, making the task more boring and potentially less challenging[42] does not change the impact of the incentive schemes. The different coefficient dimensions, in contrast to regression 1, are due to the different underlying tasks. Note that as also in BL performance is on high levels.

The Pause treatment is designed to deal with the effect of moral obligations and the introduction of opportunity costs. Figure 4 displays the regression result. Subjects solve more tasks under performance pay. However this effect is only significant for tournament pay. Performance under flat incentives is on similar levels like in BL. This is significantly higher than zero. Thus, we find evidence in favour of our intuition. The effect is, however, not caused by a drop in performance under flat payment.

Looking at the use of the pause option, the time on pause was, as expected, different from the theoretical benchmark: under flat payments subjects spent on average 20 seconds on pause which is a bit less than one third of the stage-length. Also in the non-creative task the share of participants who do not make use of the pause option is considerable (Table 8). The work productivity is not influenced by the introduction of the pause option (Table 9).[43]

The regression results for DiffPause are displayed in Figure 4. Note that performance under flat fee is about half of that in DiffBL . The number of correctly solved tasks is significantly higher with linear and tournament pay.

[41] Looking at how long subjects go on pause, the time on pause is significantly shorter under performance pay (second regression Table 6). Compared to Pause, the time that subjects spend on pause in DiffPause is significantly longer (Table 7).

[42] Table 5 displays the results of the manipulation check: subjects find the counting *1*s number indeed easier than the number adding task. However, task enjoyment and task importance are not different.

[43] Table 6 shows that subjects spent significantly less time on pause under contingent pay than under flat pay.

In this treatment, DiffPause, the work productivity under performance pay is higher than under flat payments (Table 9) and also the time on pause is lower under performance pay.[44]

4.4. 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 the overall productivity of participants in our experiment very much. 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.

With the letterset `accdeeeginst` a participant could, e.g., produce many short and simple words like `a` or `i` (1 point each), or `dan` or `ian` (6 points each). A participant could also think harder and produce longer and more complex words like `accidents` or `deceasing` (45 points). Participants, hence, face a trade off between producing either more short words or fewer long words. We take the length of the word as measure of complexity.

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 have less in common, like `denis`, `ideas`, `stance`, etc. We measure dissimilarity as the Jaro-Winkler Distance of successive words (Jaro, 1989, Winkler, 1990, van der Loo, 2014).

To measure the absolute magnitude of the effect we estimate again Equation (1), now with Y=word length and Y=word distance. 95%-confidence intervals for fixed effects are shown in Figure 5.[45] We see that incentives do have a positive impact on word length, however, only the effect of linear incentives on word length is significant.

4.5. Self-selection

4.5.1. Performance in the self-selection stages

In the last two stages of the experiment subjects select the payment scheme for the creative task and the non-creative task, respectively. To assess the change performance relative to the flat incentive in the stages with self-selection we estimate (separately for stage 7 and stage 8) the following equation:[46]

$$\text{Performance} = \beta_0 + \sum_{\text{Incent.}} \beta_{\text{inc.}} \cdot d_{\text{inc.}} + \epsilon_{\text{subj.}} \tag{2}$$

Confidence intervals are shown in Figure 6. Changes in performance are clearly different

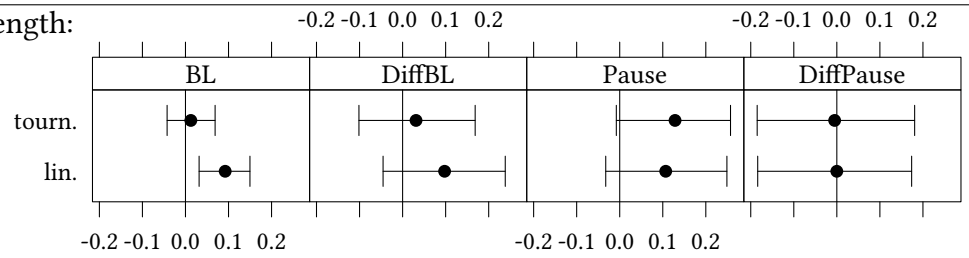
[44] The regression in Table 6 shows that in the counting *1*s task subjects spent substantially more time on pause under flat payments. Comparing the time on pause between Pause and DiffPause, Table 6 shows that in the counting *1*s task, as compared to the adding numbers task, subjects spent substantially more time on pause under flat payments. Time on pause under linear pay seems to be similar in the two treatments. Under tournament pay, subjects spent slightly less time on pause in DiffPause than in Pause. Note that the interaction effect between performance pay and DiffPause to a large extent offsets the effect of the treatment dummy.

[45] Confidence intervals for random effects are shown in Figure 12 in Appendix A.2.

[46] Since, within a single stage, there is only a single measure of performance, we do not need a random effect.

Figure 5 95%-confidence intervals for fixed effects for incentives from Equations (1) where Y is word length and word distance.

Word length:



Word distance:

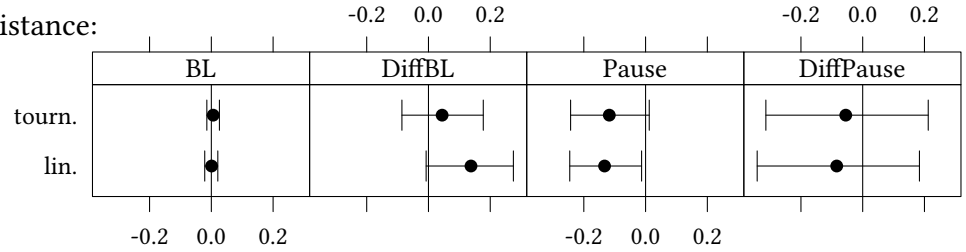
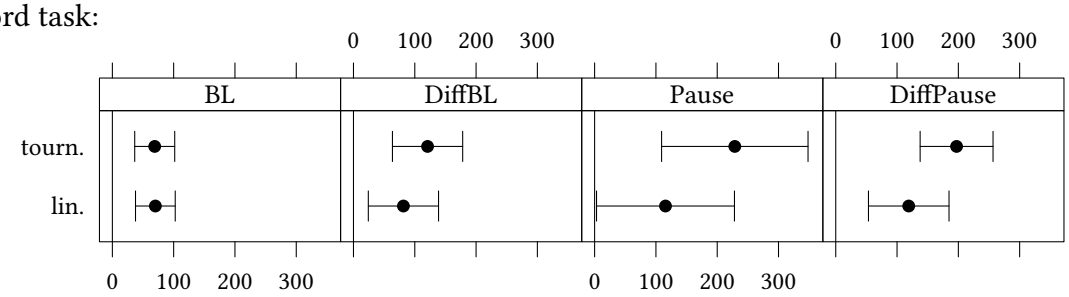
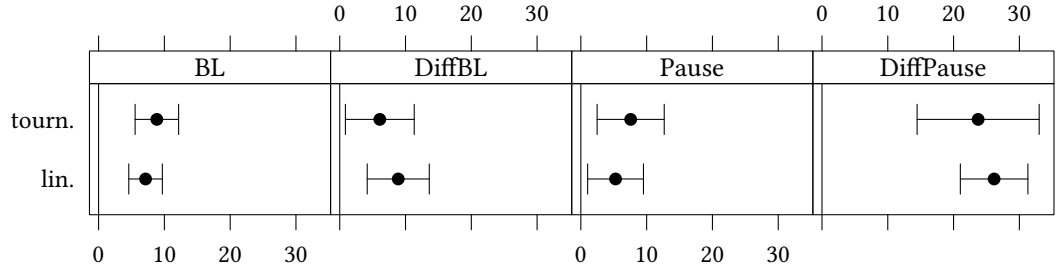


Figure 6 Incentives and performance with self selection

Word task:

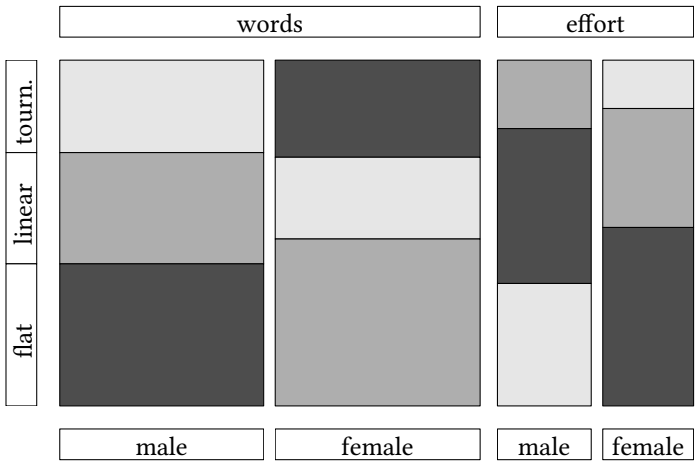


Non-creative task:



The figure shows 95%-confidence intervals for the impact of incentives (Equation 2).

Figure 7 Self-selection by Gender.



from what we found without self selction (in Figure 4). For all treatments and for both incentives we find performance to be larger with incentives.

These results seem to support hypothesis 4 for the creative and for the effort task. These results are in contrast to the results in stages 1, 3, and 5. In these stages, when the payment scheme was imposed by the experimenters, we observe effects of the payment schemes only Pause and DiffPause. Now, when the payment scheme is self-selected, those subjects who self-select into performance pay have a higher output in all treatments.

4.5.2. Frequency of selected incentives

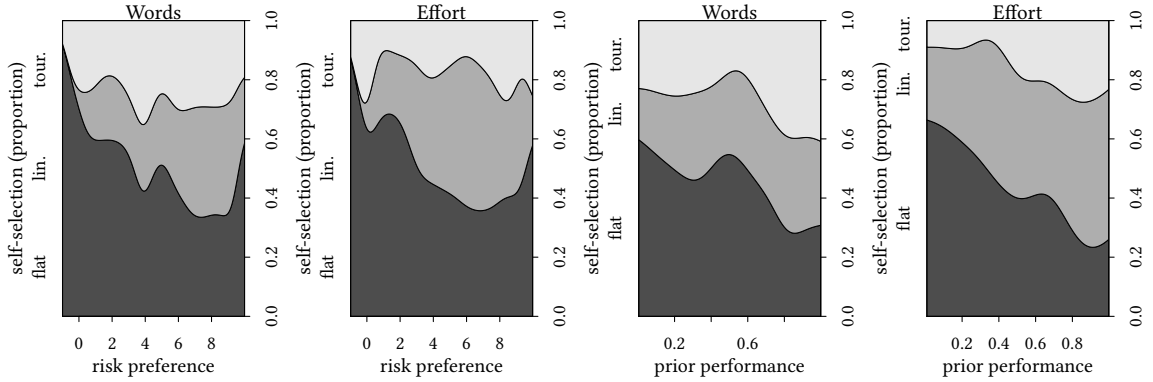
Figure 7 shows a mosaic plot how the self selection of participants into incentive schemes is affected by gender. Figure 14 in Appendix A.6 shows a mosaic plot how the self selection of participants into incentive schemes is affected by task types. In Figure 7 we see that flat incentives are chosen more frequently than, in particular, tournaments.

For our effort task we find that females are more likely to choose a flat incentive while makes are more likely to chose a linear incentive or even a tournament. This is in line with Niederle & Vesterlund (2007) who use a number adding task and find significantly more male than female participants chosing the tournament over a linear payment scheme. For the word task we find, however, only a very small difference between males and females. This is in line with Grosse & Riener (2010) who compare different task types with different gender stereotypes. For their word task they also find that men and women seem to be equally likely to choose a tournament.

4.5.3. Determinants of selected incentives

Figure 8 shows how treatment selection is determined by risk preferences and by performance in the previous task. The two panels in the left part of Figure 8 shows that the relative

Figure 8 Self-selection into treatments determined by risk and performance.



The vertical axis shows the conditional density to select into one of the three incentive schemes in stage 7 and 8 of the experiment. The horizontal axis shows for the left two panels the risk preference (as in Dohmen *et al.* (2011)). The two panels on the right show on the horizontal axis the relative performance for the first three stages of the creative task and the non-creative task, respectively. The vertical axis shows the conditional density to select into one of the three incentive schemes in stage 7 and 8 of the experiment, respectively.

frequency of choosing the flat payment decreases with more risk-loving risk preferences for the word task but also for the effort task.

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. The two panels in the right part of Figure 8 show how choosing incentives is influenced by prior performance for the word task and for the effort task. It seems that the relative frequency to choose an incentive based payment increases with higher performance in the previous stages.

To confirm what we see in the figures we estimate the following multinomial logit model:

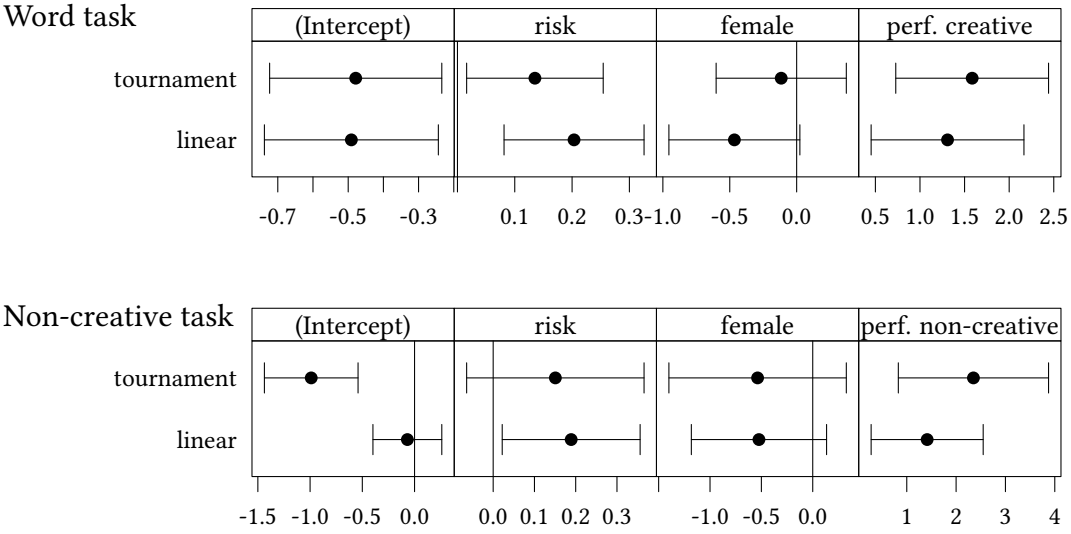
$$\log \frac{\Pr(\text{treatment})}{\Pr(\text{flat})} = \beta_{(\text{intercept})} + \beta_{\text{points}} \cdot \text{points} + \beta_{\text{risk}} \cdot \text{risk} + \beta_{\text{female}} \cdot d_{\text{female}} \qquad (3)$$

"Flat" is the reference treatment. To make the different tasks comparable, performance in the tasks is based on the percentile rank within each task. To be able to interpret intercepts as average effects we have demeaned performance, risk, and gender in the estimation of Equation 3. "Risk" is the risk measure introduced by Dohmen *et al.* (2011). Estimation results are shown in Figure 9. Estimation results for an augmented version of Equation 3 are reported in Figure 15 in the Appendix.

The coefficients for performance is clearly (and significantly) positive for all tasks and for both the linear and the tournament incentive. In line with Hypothesis 4 a good performance in the previous rounds makes it more likely to choose an incentivised treatment.

Also, more risk loving participants are more likely to select into the incentivised treatments. Again, there is no substantial difference between the effect of risk to select into the linear incentive or the tournament. Finally, there is no significant effect of gender to select in one of the incentivised treatments.

Figure 9 Determinants of treatment selection



The Figure shows 95%-confidence intervals for coefficients from Equation 3. To make the different tasks comparable, performance in the tasks is based on the percentile rank within each task. To be able to interpret the intercept, performance, risk, and gender is demeaned. Figure 15 in the Appendix presents estimation results where we also control for the treatment.

5. Discussion

Since the observed results for the creative and the non-creative tasks are qualitatively very similar, this discussion will be general. The results of this experimental study can be summed up with two main points.

First of all, subjects' behaviour in the baseline treatments (*BL*), does not differ much between the three incentive schemes (*flat*, *linear* & *tournament* pay). Also in the difficult baseline (*DiffBL*) conditions, in which the task enjoyability was modified, no effect of the incentive schemes is observed. This result suggests that task attractiveness, task enjoyment or task challenge do not seem to be the major drivers of the experimental results that we observe in the baseline treatment and in the first eperimental series.

Second, with the introduction of a paid pause option (*Pause*) we can, in particular for the creative task, observe an effect of the incentive schemes. When subjects have the possibility to make use of the paid pause option, average performance is higher under performance pay than under flat pay. This effect is driven by a lower performance under flat pay and not by a higher performance under contingent pay.[47] The effect is observable both, when we introduce the pause option alone (*Pause*), as well as when the introduction is combined with a higher task difficulty (*DiffPause*). Interestingly, within every task working *productivity* is not influenced by the availability of the pause option, with one exception: when subjects had to count *0*s and *1*s productivity is much lower when a pause option is available. Furthermore,

[47] There is one exception: in the non-creative task in the pause-treatment the positive effect of performance pay stems from a higher performance under performance-contingent pay.

none of the two performance-based incentive schemes seems to outperform the other. Still, it is striking that, even when a paid pause option is available, performance under flat pay is still significantly higher than zero.

As discussed in the introduction, with the employed experimental design we cannot distinguish whether the differences in subjects' behaviour result from the legitimacy of not working or the presence of opportunity costs. It is left for future research to disentangle these factors. Some insights into the driving factors might, however, be gotten from the questionnaire responses when subjects were asked to explain how they used the pause option. The most frequently mentioned reasons were physical ones: to take a break, to relax the eyes or because of a lack of concentration. The second most often mentioned groups of reasons are motivational and financial ones: getting a fixed amount of money and not being motivated to continue working. Besides, subjects seem to have used the pause option to receive some compensation for end-of-stage-time and to be compensated while thinking about words in the creative task. No reasons relating directly to legitimacy were given. Yet, taking a break because of physical reasons when it is officially possible is possibly be related to legitimacy.

What is striking is that in those conditions in which subjects have not only the chance, but also a financial incentive to take a break (flat in the Pause & DiffPause treatments), we again observe very high effort levels. A substantial fraction of subjects does not use the pause option at all or only for a short time. There are several potential reasons why we can observe this behaviour. First, the reasons that were discussed in the introduction (Section 1) and not targeted by the experimental design, like practicing for later periods or perceiving the experiment as an exam condition also apply here. In addition, subjects might work as this is what they came to the lab for. By signing up for the experiment they know that this time is dedicated to research and they also know that they will be compensated adequately for their participation. Potentially, the participants perceive signing up for the experiment as a kind of contract between them and the experimenter in which the pause option constitutes a test of loyalty.

Alternatively, it might be that the observed behaviour is a pure subject pool effect. All subjects participated voluntarily and perhaps people who self-selected into the experimental subjects pool are of a very hard-working type, willing to exert effort "in the name of science", no matter how they are compensated. While we cannot exclude this, subject pool effects have been discussed in other experimental settings. For example, a number of authors look at the stability of social preferences in relation to the subject pool.[48] The studies differ in their conclusions. Some studies find behaviour to be pretty similar in the lab and in the field. Often, however, the observed effects are attenuated in the field. Falk & Fehr (2003) provide a discussion of subject pool effects in the context of labour market experiments and argue that although subject pool effects are important, behaviour does not differ completely between the analysed subject pools. Thus, it is rather unlikely that our results are completely due to our specific subject pool.

An effect, which is, in the context of our experimental study, probably more relevant, is the impact of subjects' boredom in the laboratory. Even though the pause option is an outside option, it does not provide distraction. Consequently, it might be that subjects work

[48] See, for example, Falk *et al.*, 2011, Exadaktylos *et al.*, 2013 or

on the experimental tasks as this gives them something to do. By exerting effort also in those conditions where it would be payoff dominant to be on pause during the whole stage, they are indirectly willing to pay for distraction by forgoing the pause compensation. It might be a valuable idea for future research to provide subjects with an attractive outside option. Corgnet *et al.* (2013) for example developed the platform "Virtual Organization" which allows experimental subjects to easily switch between experimental (production) tasks and Internet surfing. This tool also allows to track the time of "on the job leisure". Alternatively, at the cost of losing experimental control, providing newspapers or magazines in the laboratory or allowing the subjects to work on their own work would also reduce the impact of the potentially existing boredom in the laboratory.

The last two stages of the experiment were self-selection stages. In both tasks subjects' performance is, independently of the treatment, higher under self-selected performance pay than under self-selected flat pay. Let us combine this observation with the insight that subjects with a higher number of points in the first stages of the experiment are more likely to self-select into performance pay. It seems that the observed behaviour is the result of sorting by ability (similar to the result obtained by Dohmen & Falk, 2011[49]), even though subjects do not receive relative performance feedback. Besides, gender is not significantly related to the choice of payment scheme. Dohmen & Falk have a similar finding, but in their sample risk-attitudes and gender are significantly correlated and therefore risk-attitudes capture the gender effect in their analysis. In our subject pool gender is not significantly correlated with risk attitudes However, women obtain fewer points in the first stages of the non-creative tasks. No significant relation between total number of points and gender can be found for the creative task. Hence, our results still seem to be driven by productivity sorting.

6. Conclusion

Using four different tasks, one based on creativity, one based on intelligence, and two more based on rather mindless effort, we have seen that performance depends first and foremost on individual characteristics of participants and can, on the aggregate level, hardly be influenced through financial 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 choose a performance-dependent 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

[49] In fact the number adding task is similar in nature to the task that Dohmen & Falk used, namely a math task in which subjects had to multiply numbers.

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 particularly striking that we observe only small effects of incentives in the baseline version non-creative tasks.

We vary task difficulty and outside options to better understand this phenomenon. We find that making tasks more difficult or less interesting alone does not change the results. With the introduction of opportunity costs, however, we observe differences of incentive schemes on subjects performance. This difference is caused by a lower performance under flat payment when the pause option is available. The size of the difference differs between the tasks, yet the direction is the same across all tasks and levels of difficulty (except for the non-creative task in the Pause treatment). Interestingly, while under flat pay it would have been payoff-maximising to not work at all and be on pause the whole time, subjects still exert a significant amount of effort and performance is considerable. In the self-selection stage we observe that under flat pay subjects perform less well than when they self-select themselves into performance pay.

To conclude, we were aiming at exploring potential reasons why we did not observe effects of financial incentive schemes on subjects' performance in creative and non-creative tasks. Our results suggests that task attractiveness, task enjoyment or task challenge do not seem to be the main drivers of our earlier observations. However, when subjects have the possibility to make use of an incentivised pause option, their performance stays high under performance pay and decreases under flat pay while their productivity remains unchanged in almost all conditions.

The experimental observations give some directions for potential future research. First of all, with the employed design it cannot be distinguished whether the observed effect in the Pause and difficult Pause treatments stems from the introduction of opportunity costs or from the fact that also a legitimacy of taking a break is introduced. An experimental treatment which uses a non-compensated pause option could help to shed some light on the driving factor. Observing that subjects exert substantial efforts also under flat pay, and even forego payments in the Pause and difficult Pause treatments when they do not use the pause option as often as possible, indicates that subjects must be driven by something else than pure financial payoff maximisation. Possibly, subjects work to not be bored. Thus, a potential treatment to analyse this point further would be to provide subjects with an attractive outside option to pursue on the job leisure (like e.g. (Corgnet *et al.*, 2013).

This study shows that the availability of outside attractive options can be an influential factor and it might be important to give more attention to it when designing economic experiments. In particular, considering that in "real life" outside options are often available, it is important to examine whether experimental results are robust to the availability of attractive outside options. Maybe it is particular important to keep this in mind in labour economic experiments. This study demonstrated that for creative and non-creative tasks, independent of the level of difficulty or attractiveness, results are influenced by the availability of an outside option.

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.

Bailey, C.D. & Fessler, N.J. (2011). The moderating effects of task complexity and task attractiveness on the impact of monetary incentives in repeated tasks. *Journal of Management Accounting Research*, *23*(1), pp. 189–210.

Barsh, J., Capozzi, M. & Mendonca, L. (2007). How companies approach innovation: a McKinsey global survey. *The McKinsey Quarterly*.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), pp. 1–48.

Blumkin, T., Ruffle, B. & Ganun, Y. (2010). *Are income and consumption taxes ever really equivalent? Evidence from a real-effort experiment with real goods*, Discussion Paper 5145, IZA, Forschungsinstitut zur Zukunft der Arbeit, Bonn.

Bonner, S.E., Hastie, R., Sprinkle, G.B. & Young, S.M. (2000). A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. *Journal of Management Accounting Research*, *12*(1), pp. 19–64.

Bradler, C., Neckermann, S. & Warnke, A.J. (2013). *Rewards and performance: A comparison across creative and routine tasks*, mimeo, ZEW Mannheim.

Bradler, C., Neckermann, S. & Warnke, A.J. (2019). Incentivizing creativity: A large-scale experiment with performance bonuses and gifts. *Journal of Labor Economics*, *37*(3), pp. 793–851.

Brase, G.L. (2009). How different types of participant payments alter task performance. *Judgement and Decision Making*, *4*(5), pp. 419–429.

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.

Corgnet, B., Hernan-Gonzalez, R. & Rassenti, S. (2013). *Peer pressure and moral hazard in teams: Experimental evidence*, Working paper 13-01, Chapman University, Economic Science Institute.

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.

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), pp. 627–668.

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

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*.

Dohmen, T. & Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *The American Economic Review*, *101*(2), pp. 556–590.

Duncker, K. & Lees, L. (1945). On problem-solving. *Psychological monographs*, *58*(5), pp. i–113.

Eisenberger, R., Pierce, W.D. & Cameron, J. (1999). Effects of reward on intrinsic motivation - negative, neutral, and positive: Comment on Deci, Koestner, and Ryan (1999). *Psychological Bulletin*, *125*, pp. 677–691.

Eisenberger, R. & Selbst, M. (1994). Does reward increase or decrease creativity? *Journal of Personality and Social Psychology*, *66*(6), p. 1116.

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

Exadaktylos, F., Espín, A.M. & Brañas-Garza, P. (2013). Experimental subjects are not different. *Scientific reports*, *3*.

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 Americal Economic Review*, pp. 1611–1630.

Falk, A. & Fehr, E. (2003). Why labour market experiments? *Labour Economics*, *10*(4), pp. 399–406.

Falk, A., Meier, S. & Zehnder, C. (2011). Do lab experiments misrepresent social preferences? The case of self-selected student samples. *Journal of the European Economic Association, forthcoming*.

Fessler, N.J. (2003). Experimental evidence on the links among monetary incentives, task attractiveness, and task performance. *Journal of Management Accounting Research*, *15*(1), pp. 161–176.

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.

Gamage, D., Hayashi, A. & Nakamura, B. (2010). Experimental evidence of tax framing effects on the work/leisure decision. *Berkeley Olin Program in Law & Economics, Working Paper Series*.

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., Niederle, M. & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, pp. 1049–1074.

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.

Grosse, N.D. & Riener, G. (2010). *Explaining gender differences in competitiveness: Gender-task stereotypes*, Jena Economic Research Papers 2010-017, Friedrich-Schiller-University Jena.

Guilford, J. (1950). Creativity. *American Psychologist*, *5*, pp. 444–454.

Harbring, C. & Irlenbusch, B. (2008). How many winners are good to have? On tournaments with sabotage. *Journal of Economic Behavior & Organization*, *65*(3), pp. 682–702.

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.

Hertwig, R. & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(03), pp. 383–403.

Houser, D., Schunk, D., Winter, J. & Xiao, E. (2010). *Temptation and commitment in the laboratory*, IEW - Working Papers 488, Institute for Empirical Research in Economics - University of Zurich.

Huhtala, H. & Parzefall, M.R. (2007). A review of employee well-being and innovativeness: An opportunity for a mutual benefit. *Creativity and Innovation Management*, *16*(3), pp. 299–306.

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.

Lezzi, E., Fleming, P. & Zizzo, D.J. (2015). *Does it matter which effort task you use? A comparison of four effort tasks when agents compete for a prize*, Working Paper series, University of East Anglia, Centre for Behavioural and Experimental Social Science (CBESS) 15-05, School of Economics, University of East Anglia, Norwich, UK.

Mohnen, A., Pokorny, K. & Sliwka, D. (2008). Transparency, inequity aversion, and the dynamics of peer pressure in teams: Theory and evidence. *Journal of Labor Economics*, *26*(4), pp. 693–720.

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.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of economic literature*, *37*(1), pp. 7–63.

R Development Core Team (2019). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

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), pp. 166–183.

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.

Sternberg, R.J. (1999). *Handbook of creativity*, Cambridge University Press.

Stone, T.H. (1971). Effects of mode of organization and feedback level on creative task groups. *Journal of Applied Psychology*, *55*(4), p. 324.

Torrance, E.P. (1968). *Torrance tests of creative thinking*, Personnel Press, Incorporated.

van der Loo, M. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*, pp. 111–122.

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

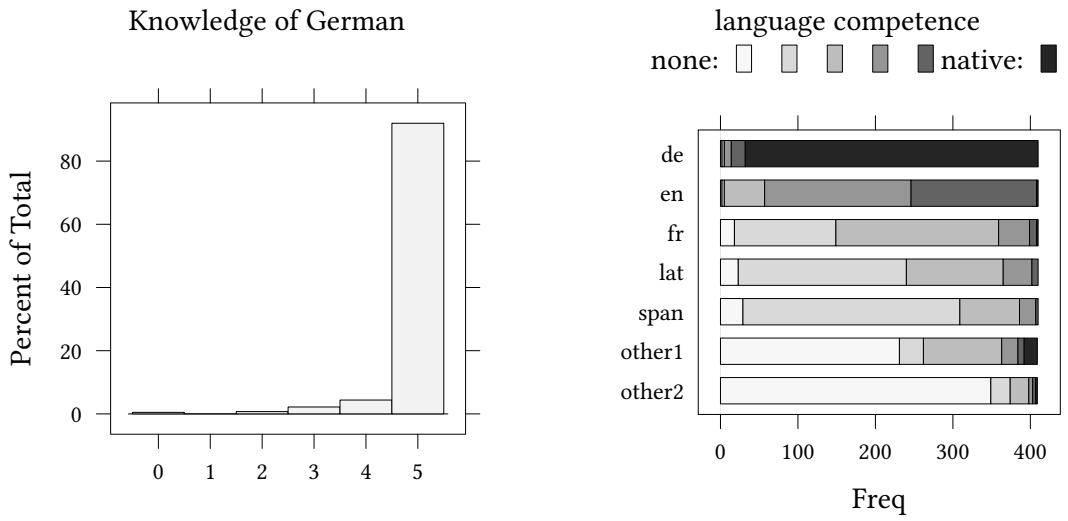
West, M.A. (2002). Sparkling fountains or stagnant ponds: An integrative model of creativity and innovation implementation in work groups. *Applied Psychology*, *51*(3), pp. 355–387.

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.

Yerkes, R.M. & Dodson, J.D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*(5), pp. 459–482.

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

Figure 10 Language competence



The figure includes all 411 participants from both experiments.

A. Appendix

A.1. Language competence

Figure 10 shows on the left the distribution of the knowledge of German for our participants on a range from 1=none to 5=native. The right graph show the competence for other languages.

A.2. Random effects for Equations (1

95%-confidence intervals for standard deviations of the random effects for Equation (1) for Y=Performance are shown in Figure 11. Since the distribution of standard deviations is not necessarily symmetric we are using a percentile bootstrap. 95%-confidence intervals for standard deviations of the random effects for Equation (1) when Y is word length and when Y is word distance are shown in Figure 12. Since the distribution of standard deviations is not necessarily symmetric we are using a percentile bootstrap. We use 500 replications using `bootMer` from `lme4` 1.1-21.

A.3. Manipulation check for "difficult" tasks

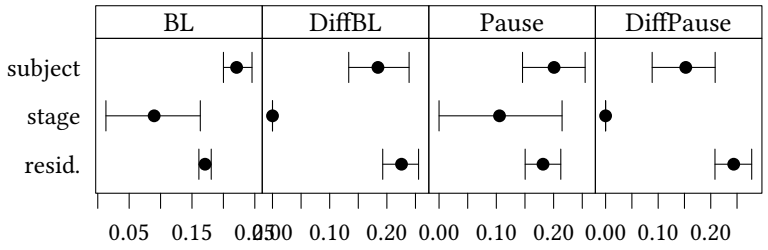
To assess whether our task manipulation works we estimate the following equation:

$$\text{Evaluation} = \beta_0 + \sum_{\text{treat.}} \beta_{\text{treat.}} \cdot d_{\text{treat.}} + \epsilon_{\text{subj.}} \tag{4}$$

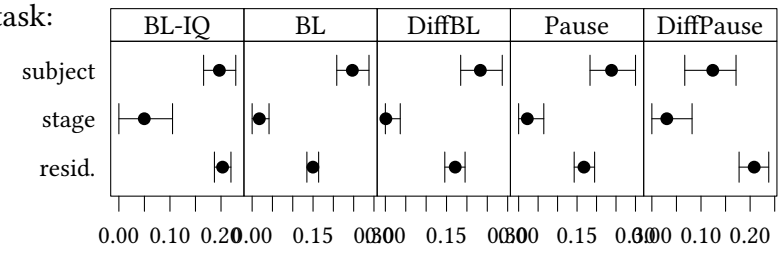
Results for the creative task are shown in Table 4. Results for the non-creative tasks are shown in Table 5.

Figure 11 Random effects for Equation (1)

Word task:



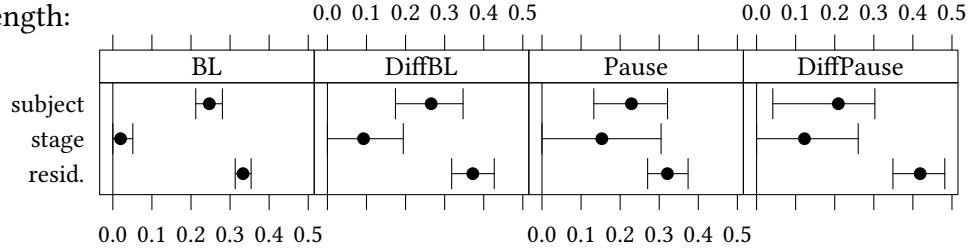
Non-creative task:



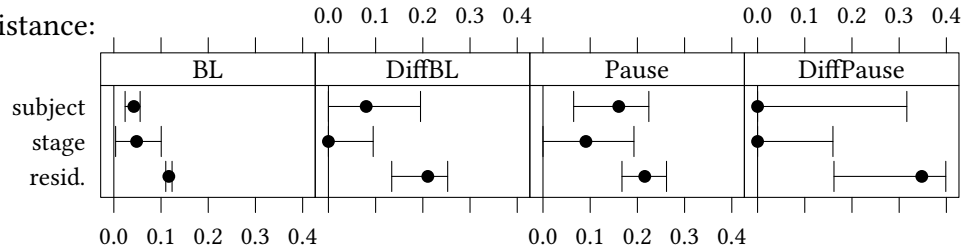
The figure shows 95%-confidence intervals for the standard deviation of random effects for subject and stage as well as the standard deviation of the residual (sigma).

Figure 12 Random effects when Y=word length and word distance, Equation (1).

Word length:



Word distance:



The figure shows 95%-confidence intervals for the standard deviation of random effects for subject and stage as well as the standard deviation of the residual (sigma).

Table 4 Evaluation of the creative tasks (Eq. 4)

	enjoy	difficult	importance
(Intercept)	4.11^{***}	2.96^{***}	6.77^{***}
	(0.14)	(0.10)	(0.37)
DiffBL	0.50	3.53^{***}	0.11
	(0.35)	(0.25)	(0.53)
DiffPause	1.68^{***}	2.89^{***}	−0.14
	(0.36)	(0.25)	(0.54)
Pause	2.19^{***}	1.78^{***}	0.11
	(0.38)	(0.26)	(0.56)
R^2	0.11	0.43	0.00
Adj. R^2	0.10	0.43	-0.01
Num. obs.	410	410	195
RMSE	2.29	1.60	2.70

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5 Evaluation of the non-creative tasks (Eq. 4)

	enjoy	difficult	importance
(Intercept)	3.95^{***}	4.11^{***}	6.52^{***}
	(0.26)	(0.20)	(0.37)
DiffBL	1.09^{*}	-2.01^{***}	1.05^{*}
	(0.45)	(0.35)	(0.53)
DiffPause	1.42^{**}	-2.13^{***}	0.58
	(0.46)	(0.36)	(0.54)
Pause	0.63	0.96^{*}	0.11
	(0.48)	(0.37)	(0.56)
R^2	0.05	0.26	0.02
Adj. R^2	0.03	0.25	0.01
Num. obs.	245	245	195
RMSE	2.64	2.06	2.69

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6 Estimation of Equation 5 for the creative task: time on pause
Estimation baseline: *flat*

	words-Pause	effort-Pause	words-DiffPause	effort-DiffPause
(Intercept)	15.77^{***}	20.35^{***}	15.77^{***}	20.35^{***}
	(1.67)	(1.84)	(1.67)	(1.84)
incentivelinear	-9.99^{***}	-18.02^{***}	-9.99^{***}	-18.02^{***}
	(1.84)	(2.43)	(1.84)	(2.43)
incentivetournament	-8.86^{***}	-17.05^{***}	-8.86^{***}	-17.05^{***}
	(1.84)	(2.43)	(1.84)	(2.43)
AIC	12017.54	12321.19	12017.54	12321.19
BIC	12048.24	12351.86	12048.24	12351.86
Log Likelihood	-6002.77	-6154.60	-6002.77	-6154.60
Num. obs.	1233	1226	1233	1226
Num. groups: subject	411	411	411	411
Num. groups: stage	3	3	3	3
Var: subject (Intercept)	453.45	175.22	453.45	175.22
Var: stage (Intercept)	0.00	0.00	0.00	0.00
Var: Residual	697.41	1201.18	697.41	1201.18

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

A.4. Time on pause

We will estimate two different mixed effects models to analyse the time that subjects spent on pause. The estimation results of Equation 5 for the creative task are displayed in Table 6. The regression is run for the two pause treatments separately. The estimation baseline is flat pay. The estimation results of Equation 6 are displayed in Table 7.

$$\text{Time on pause} = \beta_0 + \sum_{\text{inc.}} \beta_{\text{inc.}} \cdot d_{\text{inc.}} + \epsilon_{\text{stage}} + \epsilon_{\text{subj.}} + \epsilon_{\text{subj.},t} \tag{5}$$

$$\begin{aligned} \text{Time on pause} = \beta_0 &+ \sum_{\text{inc.}} \beta_{\text{inc.}} \cdot d_{\text{inc.}} + \beta_{\text{DiffPause}} \cdot d_{\text{DiffPause}} \\ &+ \sum_{\text{inc.}} \beta_{\text{DiffPause}\cdot\text{inc.}} \cdot d_{\text{DiffPause}\cdot\text{inc.}} + \epsilon_{\text{stage}} + \epsilon_{\text{subj.}} + \epsilon_{\text{subj.},t} \end{aligned} \tag{6}$$

A.5. Other statistics

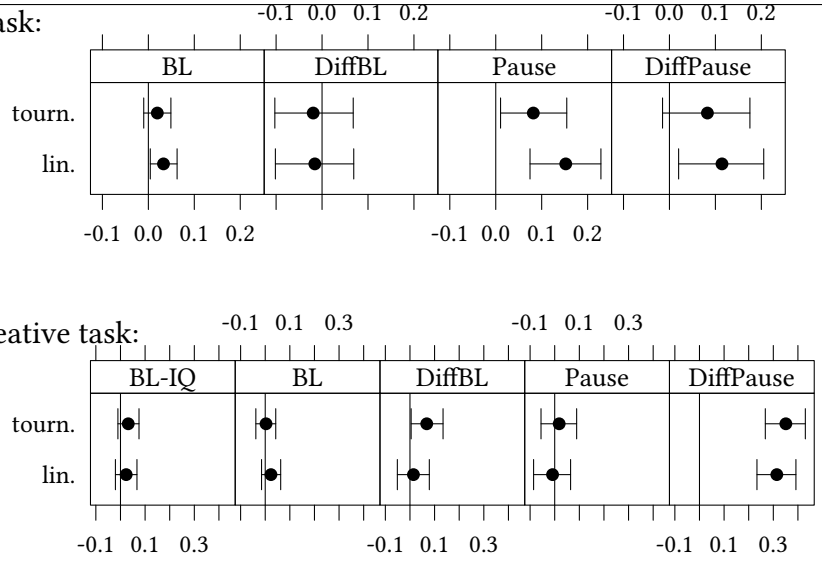
A.6. Self-selection

Figure 14 shows a mosaic plot how the self selection of participants into incentive schemes is affected by task types.

Table 7 Estimation of Equation 6 for the creative task: time on Pause
Estimation baseline: *Pause-flat*

	β	[95%,	CI]
(Intercept)	53.7	45.2	62.6
treatmentBL	-53.7	-63.3	-44.6
treatmentDiffBL	-53.7	-65.6	-42.2
treatmentDiffPause	31.5	20.5	42
incentivelinear	-37.6	-48.1	-27.2
incentivetournament	-27.7	-38.8	-17.7
treatmentBL:incentivelinear	37.6	26.8	48.8
treatmentDiffBL:incentivelinear	37.6	23.4	51.6
treatmentDiffPause:incentivelinear	-13.3	-26.5	0.648
treatmentBL:incentivetournament	27.7	16.6	39.8
treatmentDiffBL:incentivetournament	27.7	13.9	42.3
treatmentDiffPause:incentivetournament	-22.4	-35.4	-7.97

Figure 13 95%-confidence intervals for fixed effects for incentives from Equation (1).



This is a variant of Figure 4 where performance is not represented as the percentile rank in the respective task. Instead performance is measured as the number of "points" participants obtained in the experiment.

Table 8 Proportion (in %) of participants who do not use the pause option

	Incentive	β	[95%	CI]
creatives task, Pause	flat	82.24	78.19	85.81
	linear	85.40	81.61	88.67
	tournament	83.70	79.77	87.14
effort task, Pause	flat	85.43	81.61	88.72
	linear	93.19	90.30	95.43
	tournament	90.73	87.50	93.36
creatives task, DiffPause	flat	82.24	78.19	85.81
	linear	85.40	81.61	88.67
	tournament	83.70	79.77	87.14
effort task, DiffPause	flat	85.43	81.61	88.72
	linear	93.19	90.30	95.43
	tournament	90.73	87.50	93.36

β represents the share of participants who do not use the pause option. Confidence intervals are exact.

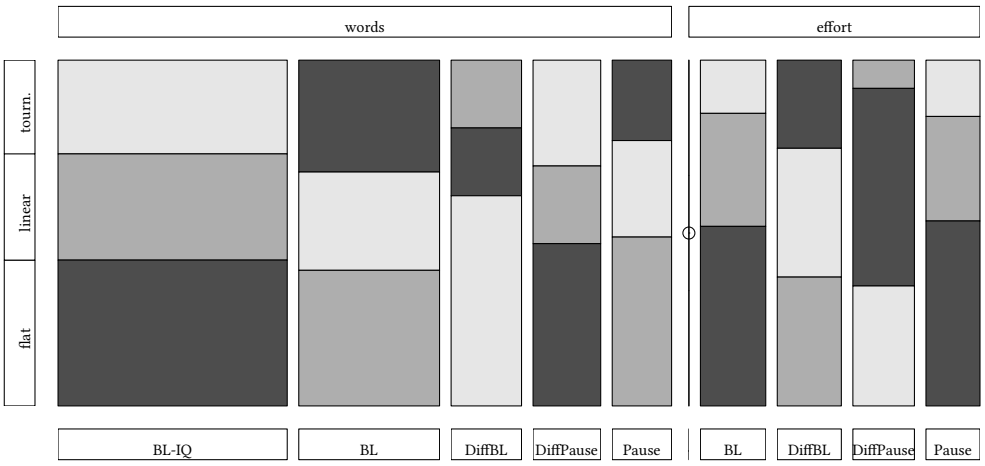
Table 9 Estimation of Equation 1 for *productivity* in the non-creative task
Estimation baseline: *flat*

	words-Pause	effort-Pause	words-DiffPause	effort-DiffPause
(Intercept)	0.86***	0.04***	0.86***	0.04***
	(0.05)	(0.00)	(0.05)	(0.00)
incentivelinear	0.05*	0.00**	0.05*	0.00**
	(0.02)	(0.00)	(0.02)	(0.00)
incentivetournament	0.03	0.00***	0.03	0.00***
	(0.02)	(0.00)	(0.02)	(0.00)
AIC	916.24	-5788.18	916.24	-5788.18
BIC	946.95	-5757.51	946.95	-5757.51
Log Likelihood	-452.12	2900.09	-452.12	2900.09
Num. obs.	1233	1226	1233	1226
Num. groups: subject	411	411	411	411
Num. groups: stage	3	3	3	3
Var: subject (Intercept)	0.09	0.00	0.09	0.00
Var: stage (Intercept)	0.01	0.00	0.01	0.00
Var: Residual	0.07	0.00	0.07	0.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Productivity, measured in points per second, is calculated as the points acquired in one stage, divided by the working time of the individual.

Figure 14 Self-selection by Treatment.



Equation 7 is an extended version of Equation 3:

$$\log \frac{\Pr(\text{incentive})}{\Pr(\text{flat})} = \beta_{(\text{intercept})} + \beta_{\text{treat.}} \cdot d_{\text{treatment}} + \beta_{\text{points}} \cdot \text{points} + \beta_{\text{risk}} \cdot \text{risk} + \beta_{\text{fem.}} \cdot d_{\text{female}} + \beta_{\text{fem.·treat.}} \cdot d_{\text{female}} \cdot d_{\text{treat.}} \tag{7}$$

Estimation results are shown in Figure 15.

A.7. Lettersets

A.7.1. A British 75%-quantile letterset

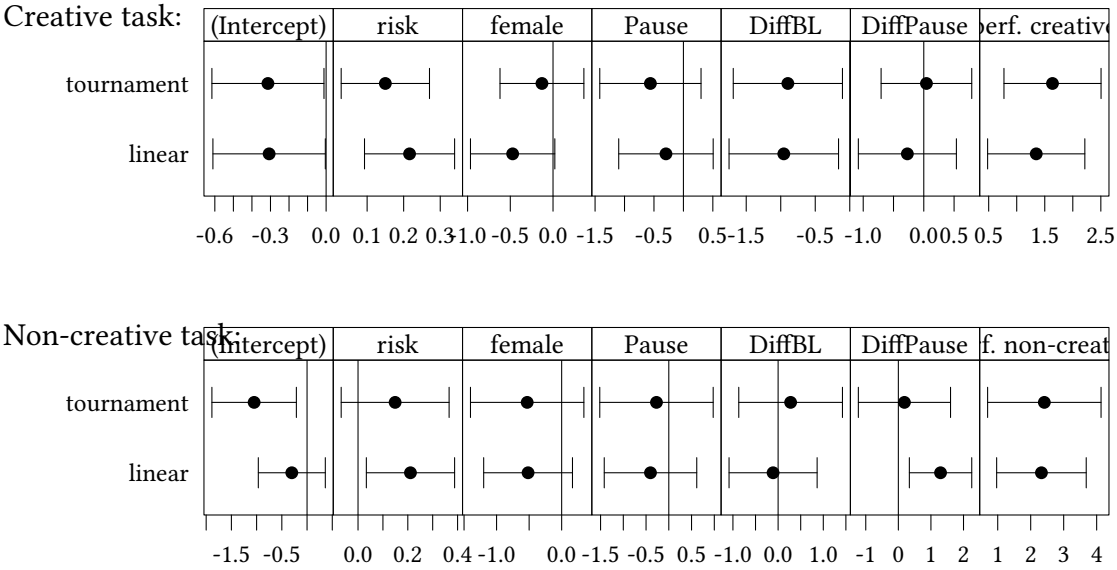
This letterset is similar to the German lettersets that we used in the experiment. The only difference is that it has been built with the British ispell dictionary.

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[50] (here 0.888156). We restricted our attention to lettersets which were close (within 1% margin) to the 75% quantile for achievable 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

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

Figure 15 Multinomial logit for incentive selection in stages 7 and 8, equation 7



To make the different tasks comparable, performance in the tasks is based on the percentile rank within the given task. To be able to interpret the intercept, performance, risk, and gender is demeaned.
This is an extended version of the model presented in Figure 9.

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 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

Table 10 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

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

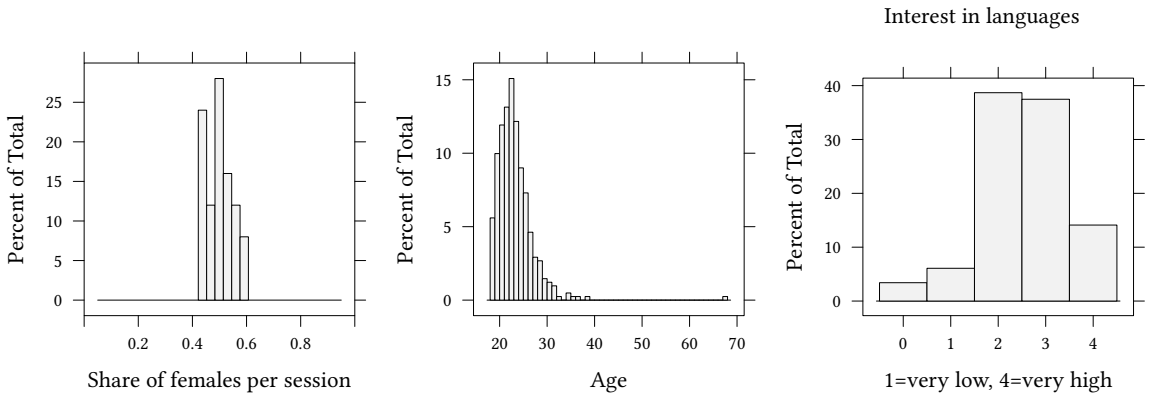
A.7.2. A German 75%-quantile letterset

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

Figure 16 Hobbies and interest in languages



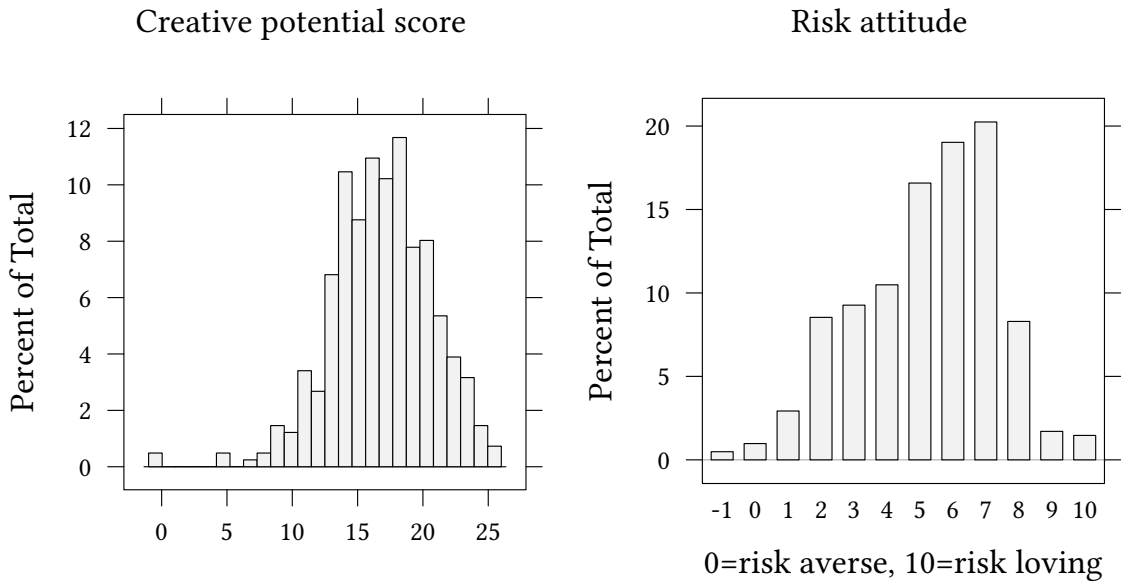
A.8. Subject pool

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 16). 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 17.

Risk-preferences were elicited as in (Dohmen *et al.*, 2011). The 11-point scale reaches from 0 (very risk-averse) to 10 (very risk-loving). The distribution is shown in Figure 17.

Figure 17 shows participants' risk preferences according to Dohmen *et al.* (2011).

Figure 17 Creativity and attitude toward risk



The Creative potential score was elicited as in (DiLiello & Houghton, 2008). Risk-preferences were elicited as in (Dohmen *et al.*, 2011).