The Effect of Incentives in Non-Routine Analytical Team Tasks - Evidence From a Field Experiment

Florian Englmaier (LMU Munich)
Stefan Grimm (LMU Munich)
David Schindler (Tilburg University)
Simeon Schudy (LMU Munich)

Discussion Paper No. 71

February 8, 2018

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Dr. Klaus M. Schmidt, University of Munich, 80539 Munich, Germany
+49 (89) 2180 3405 | info@rationality-and-competition.de

The Effect of Incentives in Non-Routine Analytical Team Tasks—Evidence from a Field Experiment*

Florian Englmaier† Stefan Grimm‡ David Schindler§
Simeon Schudy¶

February 22, 2018

Abstract

Despite the prevalence of non-routine analytical team tasks in modern economies, little is known about how incentives influence performance in these tasks. In a field experiment with more than 3000 participants, we document a positive effect of bonus incentives on the probability of completion of such a task. Bonus incentives increase performance due to the reward rather than the reference point (performance threshold) they provide. The framing of bonuses (as gains or losses) plays a minor role. Incentives improve performance also in an additional sample of presumably less motivated workers. However, incentives reduce these workers' willingness to "explore" original solutions.

JEL codes: C92, C93, J33, D03, M52

Keywords: team work, bonus, incentives, loss, gain, non-routine, exploration

*We thank Steffen Altmann, Oriana Bandiera, Iwan Barankay, Erlend Berg, Jordi Blanes i Vidal, Alexander Cappelen, Eszter Czibor, Robert Dur, Florian Ederer, Constança Esteves-Sorenson, Armin Falk, Urs Fischbacher, Guido Friebel, Holger Herz, David Huffman, Lorenz Götte, Michael Kosfeld, Botond Kőszegi, Andreas Leibbrandt, Stephen Leider, Steven Levitt, Stephan Meier, Takeshi Murooka, Susanne Neckermann, Simon Jäger, Rajshri Jayaraman, Dirk Sliwka, Christian Traxler, Bertil Tungodden, Timo Vogelsang, Roberto Weber, as well as seminar participants at Augsburg, Barcelona, Bonn, Budapest, Columbia, Heidelberg, Johns Hopkins, Lausanne, Regensburg, Munich, Wharton, and at numerous conferences for very helpful comments. We thank Lukas Abt, Michael Hofmann, Nicolas Wuthenow, Julian Angermaier, Dominik Grothe, Katharina Hartinger, Julia Rose, Timm Opitz, Regina Seibel, Christian Boxhammer, Florian Dendorfer and Marline Wethkamp for excellent research assistance. Stefan Grimm acknowledges funding by the German Research Foundation (DFG) through GRK 1928. Financial support by the DFG through CRC TRR 190 is also gratefully acknowledged. This study was approved by the Department of Economics' Institutional Review Board (IRB) at the University of Munich (Project 2015-11).

†florian.englmaier@econ.lmu.de, +49 89 2180 5642, University of Munich, Department of Economics & Organizations Research Group (ORG), Geschwister-Scholl-Platz 1, D-80539 Munich, Germany.

‡stefan.grimm@econ.lmu.de, +49 89 2180 9787, University of Munich, Department of Economics, Geschwister-Scholl-Platz 1, D-80539 Munich, Germany.

§d.schindler@uvt.nl, +31 13 466 4838 Tilburg University, Department of Economics, PO Box 90153, 5000 LE Tilburg, Netherlands.

¶simeon.schudy@econ.lmu.de, +49 89 2180 9786, University of Munich, Department of Economics, Geschwister-Scholl-Platz 1, D-80539 Munich, Germany.

1 Introduction

Until the 1970s, a major share of the workforce performed predominantly manual and repetitive routine tasks with little need to coordinate in teams. Since then, we have witnessed a rapidly changing work environment. Nowadays, work is frequently organized in teams (see, e.g., Bandiera et al., 2013) and a large share of the workforce performs tasks that require much more cognitive effort rather than physical labor. Autor et al. (2003) analyze task input in the US economy using four broad task categories: routine manual tasks (e.g. sorting or repetitive assembly), routine analytical and interactive tasks (e.g. repetitive customer service), non-routine manual tasks (e.g. truck driving) and non-routine analytical and interpersonal tasks (e.g. forming and testing hypotheses) and document a strong increase in non-routine analytical and interpersonal tasks between 1970 and 2000. Autor and Price (2013) reaffirm the importance of these tasks in later years.

One main feature of non-routine analytical tasks is that they confront work teams with complex and previously unknown problems. Teams are supposed to come up with innovative solutions and, in order to succeed, they need to build up and recombine knowledge (Nelson and Winter, 1982). Examples range from teams of innovative product developers to management consultant teams who have to gather, evaluate, and recombine information about their clients' problems. While this idea of recombinant innovation goes back at least to Schumpeter (1934) and has been formalized in growth theory as "recombinant growth" by Weitzman (1998), it is also central in management research. The concept of the recombination of ideas is at the core of the study of innovation, and research has repeatedly found evidence for various forms of recombination as the main mechanism producing breakthroughs; see, e.g., Fleming (2001), Hall et al. (2001), Rosenkopf and Nerkar (2001), or Gittelman and Kogut (2003).

Given the pervasiveness of these tasks in modern economies and their importance for innovation and growth, understanding the determinants of performance in these tasks is crucial. One core question is how incentives affect teams working on these cognitively demanding, interactive and diverse tasks. In many modern work environments, contracts specify performance-related bonus payments as an important part of compensation. While there is well-identified evidence about the behavioral effects of monetary incentives on performance in mechanical and repetitive routine tasks such as fruit picking, tea plucking, tree planting, sales, or production (see, e.g., Erev et al., 1993; La-

zear, 2000; Bandiera et al., 2005, 2013; Shearer, 2004; Hossain and List, 2012; Delfgaauw et al., 2015; Jayaraman et al., 2016; Englmaier et al., 2017; Friebel et al., 2017), evidence on the effects of bonus incentives is lacking for non-routine analytical tasks in which teams jointly solve a complex problem.

In this paper, we exploit a unique field setting to measure the incentive effects for joint team performance in a non-routine analytical task. We study the performance of teams in a real-life escape game in which teams have to solve a series of cognitively demanding tasks in order to succeed (usually by escaping a room within a given time limit using a key or a numeric code). These games provide an excellent setting to study non-routine analytical and interactive team tasks: teams face complex and novel problems, have to solve analytical and cognitively demanding tasks, need to collect and recombine information which requires thinking outside the box. The task is also interactive, since members of each team have to collaborate with each other, discuss possible actions, and develop ideas jointly. At the same time, real life escape games allow for an objective measurement of joint team performance (time spent until completion), as well as for exogenous variation in incentives for a large number of teams. Our particular setting allows us to vary the incentive structure for more than 900 teams in all (with more than 4,000 participants) under otherwise equal conditions and thus enables us to isolate how bonus incentives affect team performance.

Whether bonus incentives positively affect performance in such tasks is an open question as the production technology as well as the selection of workers performing such tasks may differ. Compared to mechanical and routine tasks, non-routine analytical and interactive tasks require more information acquisition, information recombination, and creative thinking. There is thus room for incentives to discourage the exploration of new and original approaches (e.g. Amabile, 1996; McCullers, 1978; McGraw, 1978; Azoulay et al., 2011; Ederer and Manso, 2013).[1] Further, non-routine analytical tasks are more likely to be performed by people who are intrinsically motivated (see, e.g., Autor and Handel, 2013; Friebel and Giannetti, 2009; Delfgaauw and Dur, 2010). In turn, extrinsic incentives could negatively affect team performance by crowding out such intrinsic motivation (e.g. Deci et al., 1999; Hennessey and Amabile, 2010; Eckartz et al., 2012; Gerhart and Fang, 2015).

[1]Takahashi et al. (2016) further argue that incentive effects may also depend on whether the task is perceived as interesting.

Recent evidence from related strands of the literature on incentives for idea creation (Gibbs et al., 2017) and creativity (e.g. Gibbs et al., 2017; Ramm et al., 2013; Bradler et al., 2014; Charness and Grieco, 2014; Laske and Schroeder, 2016), however, do not indicate negative, but mostly positive incentive effects. While these studies provide interesting insights into how certain types of incentives can affect idea creation and creative performance, they almost exclusively measure individual production, instead of team production (i.e. workers may face team incentives but work on individual tasks).[2] One rare exception is the small scale laboratory experiment by Ramm et al. (2013), which investigates the effects of incentives on the performance of two paired individuals in a creative insight problem, in which the subjects are supposed to solve the candle problem of Duncker (1945). The study find no effects of tournament incentives on performance in pairs but it is unclear whether this effect is robust, as the authors achieve rather low statistical power.

Our unique field setting allows us to substantially advance the literature on incentives for non-routine tasks. We can study the causal effect of incentives on team performance as well as on teams' willingness to explore original solutions in a non-routine analytical team task in two very distinct samples. First, we conduct a series of field experiments with regular teams (customers of our cooperation partner) who are unaware of taking part in an experiment.[3] These teams had self-selected into the task and were intrinsically motivated to solve it. Second, we investigate whether our main treatment effects are also observed in a sample of student participants in which the teams did not self-select into the task and were exogenously formed.[4] Further, by using survey responses from the student participants, we provide some initial tentative insights on how incentives affect team organization.

[2] Bradler et al. (2014), Charness and Grieco (2014) and Laske and Schroeder (2016) study individual production. In Gibbs et al. (2017), team production is potentially possible but submitted ideas have fewer than two authors on average. Similarly, recent studies on the effectiveness of incentives for teachers (Fryer et al., 2012; Muralidharan and Sundararaman, 2011), who perform at least to some extent a non-routine task, find positive effects of performance incentives but it remains unclear if and to what extent complementarities in individual teacher performance may be regarded as features of joint team production.

[3] Harrison and List (2004) classify this approach as a "natural field experiment". The study was approved by the Department of Economics' IRB at LMU Munich (Project 2015-11) and excluded customer teams with minors. Customers gave written consent that their data was to be shared with third parties for research purposes.

[4] According to Harrison and List (2004), the student sample can be considered a framed field experiment as students are non-standard subjects in the context of real life escape games.

To identify the effect of providing incentives, we implemented a between-subjects design, in which teams were randomly allocated to either a treatment condition or a control condition. For the main treatment, we offered a team bonus if the team completed the task within 45 minutes (the regular pre-specified upper limit for completing the task was 60 minutes). In the control condition, no incentives were provided. In both samples, we find that bonus incentives significantly and substantially increased performance in an objectively quantifiable dimension. Teams in the incentive treatment were more than twice as likely to complete the task within 45 minutes. Moreover, bonus incentives did not only have a local effect around the threshold for receiving the bonus but improved the performance over a significant part of the distribution of finishing times.

We leverage the advantages of our setting to study in depth the most important aspects of the incentive scheme for generating the treatment effect. We implemented the bonus incentive framed either as a gain or a loss, and find no significant differences in performance between these conditions. In contrast to earlier findings on bonus incentives for individually performed tasks (e.g., by Hossain and List, 2012; Fryer et al., 2012), our results suggest that framing might play a smaller role in non-routine, jointly solved team tasks. In addition, we implemented two treatments in the customer sample that allow us to disentangle whether bonus incentives are effective due to the performance threshold (the reference point) or the reward provided. A treatment in which we made the bonus threshold (i.e., 45 minutes) a salient reference point without providing incentives did not affect performance, whereas paying a bonus for completing the task in the regular pre-specified time of 60 minutes had a significant positive effect. Hence, the reward component seems to be key to bringing about the positive treatment effect, as opposed to merely a salient reference performance.

In order to understand what moderates the main treatment effects, we study different possible channels. Answers to our ex-post survey of the student sample suggest that incentives affect team organization in the sense that they promote the emergence of leadership and lead to a more focused and coordinated approach to solving the problem. Second, our findings (for the customer teams, who self-selected into the task) highlight that introducing incentives does not lead to a strong reduction in a team's willingness to explore innovative solutions. However, such discouragement is apparent among student teams, which were exogenously assigned to the task.

Our results provide important insights for researchers as well as practitioners in charge of designing incentive schemes for non-routine analytical team tasks. In particular, we speak to the pressing question of many practitioners, whether monetary incentives impair team performance in tasks that are non-routine and require creative thinking. This idea has recently been strongly promoted in the public, for instance by the best selling author Pink, in his famous TED talk with more than 19 million views and his popular book *Drive* (Pink, 2009, 2011). Our results alleviate most of these concerns, since we provide novel and robust evidence that bonus incentives are a viable instrument to increase performance in such tasks. The incentives in our experiment did not reduce performance but instead affected teams' outcomes positively across two distinct samples. Second, we show that it was indeed the reward component of the bonus, and not the reference point of good performance which improved teams' outcomes. The latter findings complement recent research on non-monetary means of increasing performance, in particular research referring to workers' awareness of relative performance (for a review of this literature see Levitt and Neckermann, 2014). Third, we add novel and interesting insights to the discussion of whether incentives discourage the exploration of new approaches. The answer to this question hinges crucially on the characteristics of the underlying sample. We observe such discouragement only among the student sample, in which, presumably, less intrinsically motivated teams work on the task. This result substantially extends recent laboratory findings by Ederer and Manso (2013), who show that pay-for-performance schemes can discourage the exploration of new approaches, as it informs us about when and how incentives may result in unintended consequences. Finally, we discover a novel and interesting potential channel through which incentives may improve team performance as student teams facing incentives tended to be more likely to express a desire for leadership and to report being better led.

The rest of this paper is organized as follows: Section 2 presents the field setting and the experimental design. Section 3 provides the results from both experiments. We provide a discussion in Section 4 and Section 5 concludes.

2 Experimental Design

2.1 The Field Setting

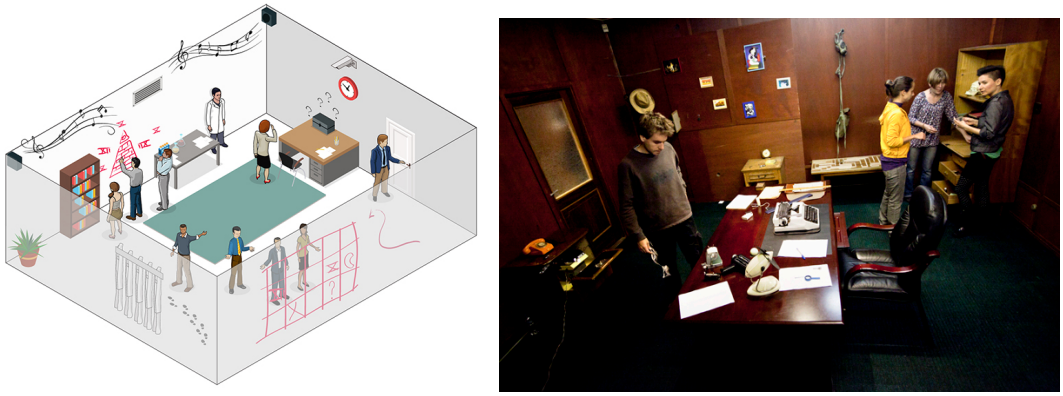
We cooperate with the company *ExitTheRoom*[5] (ETR), a provider of real-life escape games. In these games, teams of players have to solve, in a real setting, a series of tasks that are cognitively demanding, non-routine, and interactive, in order to succeed (usually by escaping from a room within a given time limit). Real-life escape games have become increasingly popular over the last years, and can now be found in almost all major cities around the globe. Often, the task is embedded in a story (e.g., to find a cure for a disease or to defuse a bomb), which is also reflected in the design of the room and how the information is presented. The task itself consists of a series of quests in which teams have to find cues, combine information, and think outside the box. They make unusual use of objects, and they exchange and develop innovative and creative ideas to solve the task they are facing within a given time limit. If a team manages to solve the task before the allotted time (one hour) expires, they win—if time runs out before the team solves all quests, the team loses.

Figure 1 ilustrates the idea and the setup of such escape rooms and shows an actual example from a real-life escape game room. The left panel is an illustration of a typical room, which contains several items, such as desks, shelves, telephones, books, and so on. These items may contain information needed to eventually solve the task. Typically, not all items will contain helpful information, and part of the task is determining which item are useful for solving the quests. The right panel shows a picture of participants actively trying to escape from their room. They already have opened drawers and closets to collect potential clues, and now jointly sort, process, and deliberate on how to use the retrieved information.

To illustrate a typical quest in a real-life escape game, we provide a fictitious example.[6] Suppose the participants have found and opened a locked box that contains a megaphone. Apart from being used as a speaker, the megaphone can also play three distinct types of alarm sounds. Among the many other items in the room, there is a volume unit (VU) meter in one corner of the room. To open a padlock on a box containing additional information, the participants will need a three digit code. The solution to this quest is to

[5] See `https://www.exittheroom.de/munich`.

[6] Our partner *ExitTheRoom* asked us to not present an actual example from their rooms.



The left panel shows typical layout of such a room, including items that might provide clues needed for a successful escape. Source: `http://www.marketwatch.com/story/the-weird-new-world-of-escape-room-businesses-2015-07-20`. The right panel shows a picture of participants actively searching their room for hints and combining the discovered information. Source: `http://boredinvancouver.com/listing/escape-game-room-experience-vancouver/`.

Figure 1: Examples of real-life escape games

play the three types of alarms on the megaphone and write down the corresponding readings from the VU meter to obtain the correct combination for the padlock. The teams at ETR solve quests similar to this fictitious example. The tasks at ETR may further include finding hidden information in pictures, constructing a flashlight out of several parts, or identifying and solving rebus (word picture) puzzles (see, also Kachelmaier et al., 2008; Erat and Gneezy, 2016).

We conducted our experiments at the facilities of *ExitTheRoom* in Munich. The location offers three rooms with different themes and background stories.[7] Teams face a time limit of 60 minutes and can see the remaining time on a large screen in their room. A room will be declared as solved if the team manages to escape from the room (or defuse the bomb) within 60 minutes. If a team does not manage to do so within 60 minutes the task is declared unsolved and the game ends. If a team gets stuck, they can request hints via radio from the staff at ETR. As they can only ask for up to five hints in all, a team needs to state explicitly that they want to receive a hint. The hints never state the direct solution to a task, but only provide vague clues regarding the next required step.

[7] *Zombie Apocalypse* requires teams to find the correct mix of liquids before time runs out (the anti-Zombie potion), in *The Bomb*, a bomb and a code to defuse it has to be found, and in *Madness*, teams need to find the correct code to open a door so as to escape (ironically) before a mad researcher experiments on them. For the sake of the reader, in the main text we refrain from presenting the regression specifications with room fixed effects. We provide these specifications in the Appendix. Adding room fixed effects does not change our results (see Table A.9).

The setting at ETR reflects many aspects of modern non-routine analytical team tasks. First, finding clues and information very much matches the activity of research that is often necessary before collaborative team work begins. Second, combining the discovered information is not trivial, and requires ability for creative problem solving. The subjects are required to process stimuli in a way that transcends the usual thinking patterns, or are required to make use of objects in unusual ways. Third, to solve the task, the subjects must effectively cooperate as a team. As in actual work environments, where the individuals in a team are supposed to provide additional angles on the problem at hand, different approaches to problem solving will enable a team to solve the task more quickly. Lastly, participants who self-select into the task have a strong motivation to succeed as they have spent a non-negligible amount of money to perform the task (participants pay between €79 (for two-person groups) and €119 (for six-person groups) for a one-hour game). We interpret the fact that many teams opt to write their names and finishing times on the walls of the entrance area of ETR as evidence for such a strong motivation. Another, more objective, reason to solve the task quickly is the fact that at any given point in time, teams do not know how many quests are left to solve the task in its entirety. That is, if a team wants to succeed, they have an incentive to succeed quickly.

While these features provide an excellent framework for studying the effect of incentives on team performance, the setting is also extremely flexible. The collaboration with ETR allows implementing different incentives for more than 700 teams of customers and studying whether incentives increase performance also in a sample of presumably less motivated and exogenously formed teams of student participants. In particular, it affords a unique opportunity to compare incentive effects for teams who have self-selected into the task (regular customers) and incentive effects for teams who were confronted with the task by us, i.e., teams who perform the task as part of their paid participation in an economic experiment.

2.2 Experimental Treatments and Measures of Performance

We conducted the field experiment with 3308 customers of *ExitTheRoom* Munich and implemented a between-subjects design. Our main treatments included 487 teams who were randomly allocated to either the control condition or a bonus incentive condition. In the bonus condition, *Bonus45* (249 teams), a team received a monetary bonus for the

team if they managed to solve the task in less than 45 minutes. In the *Control* condition (238 teams), teams were not offered any bonus. We framed the bonus either as a gain (125 teams) or as a loss (124 teams). In *Gain45*, each team was informed that they would receive the bonus if they managed to solve the task in less than 45 minutes. In *Loss45*, each team received the bonus in cash up front, kept it during their time in the room, and were informed that they would have to return the money if they did not manage to solve the task in less than 45 minutes.[8]

Additionally, we ran two experimental treatments that allow us to test whether bonus incentives were effective because of the monetary benefits or because the 45-minute threshold worked as a salient reference point. In the first additional treatment (*Reference Point*, 147 customer teams), we explicitly mentioned the 45 minutes as a salient reference point before the team started working on the task, but did not pay any bonus. We said: "In order for you to judge what constitutes a good performance in terms of remaining time: if you make it in 45 minutes or less, that is a very good result." In treatments *Gain60* (42 customer teams) and *Loss60* (46 customer teams), we provided a monetary bonus but did not provide the reference point of 45 minutes: teams received the bonus if they solved the task within 60 minutes.

We collected observable information related to team performance and team characteristics, which include time needed to complete the task, number and timing of requested hints, team size, gender and age composition of the team[9], team language (German or English), experience with escape games[10], and whether the customers came as a private group or were part of a company team building event. Our primary outcome variable

[8]The bonus amounted, on average, to approximately €10 per team member. Teams in the field experiments received a bonus of €50 (for the entire team of between two and eight members, on average about five). To keep the per-person incentives constant in the student sample with three team members (described below), the student teams received a bonus of €30. The treatment intervention (i.e. the bonus announcement) was always implemented by the experimenter present on site. For that purpose, he or she announced the possibility of the team's earning a bonus and had the teams sign a form (see Appendix A.2) indicating that they understood the conditions for receiving (in *Gain45*) or keeping (in *Loss45*) the bonus. The bonus incentive was described as a special offer and no team questioned that statement. The experimenter also collected the data. We always made sure that the experimenters blended in with the ETR staff.

[9]In order to preserve the natural field experiment, we did not interfere with the usual procedures of ETR. Thus we did not explicitly elicit participants' ages. Instead, the age of each participant was estimated based on appearance to be either 1) below 18 years, 2) between 18 and 25 years, 3) between 26 and 35 years, 4) between 36 and 50 years, 5) 51 years or older. Teams with members estimated to be minors were excluded from the experiment (following the request by the IRB).

[10]ETR staff ask teams whether they have ever participated in an escape game irrespective of our experiment.

is team performance, which we measure by i) whether or not teams solved the task in 45 minutes and by ii) the time left upon completing the task. Comparing the incentive treatments with the control condition allows us to estimate the causal effect of bonus incentives on these objective performance measures. The difference between performance in *Loss45* and *Gain45* allows us to determine whether there is an additional benefit from providing incentives in a loss frame. Differences in performance between *Reference Point* and *Control* reveal whether the reference point of 45 minutes increased the performance of the teams even if a monetary bonus was absent. The performance in *Gain60* and *Loss60* as compared to *Control* allows an additional test of whether the monetary component of the bonus was effective even when there was no change in the reference point as compared to the control.[11]

Further, we replicated our main treatments (*Gain45*, *Loss45* and *Control*) in a framed field experiment at *ExitTheRoom* in which we randomly allocated student participants from the subject pool of the social sciences laboratory at the University of Munich (MELESSA) to teams (804 participants in 268 teams). The additional sample allows us to study whether bonuses affect team performance in similar ways when the team composition was exogenous and the teams did not themselves choose to perform the non-routine task. Further, it enables us to collect additional data on task perception and team organization.

2.3 Procedures

2.3.1 Natural Field Experiment (Customer Sample)

We conducted the field experiment with customers of *ExitTheRoom* during their regular opening hours from Monday to Friday.[12] We implemented the main treatments of the field experiment (*Gain45*, *Loss45* and *Control*) in November and December 2015 and from January to May 2017. In the second phase of data collection we further ran the additional treatments *Loss60*, *Gain60* and *Reference Point*. To ensure that each room entered each condition with a similar frequency, we randomized on a daily level. This allowed us to also to avoid treatment spillovers among different teams on site (as participants from one

[11] Note that in *Control*, roughly 10 percent of the teams solved the task within 45 minutes, whereas roughly 70 percent did so within 60 minutes. Hence, the treatments which paid a bonus for solving the task in 60 minutes reveal also whether bonuses worked even if they did not refer to extraordinary performance.

[12] *ExitTheRoom* offers time slots from Monday through Friday from 3:45 p.m. to 9:45 p.m., and Saturday and Sunday from 11:15 a.m. to 9:45 p.m., with the different rooms shifted by 15 minutes to avoid overlaps and congregations of teams in the hallway.

slot could potentially encounter participants arriving early for the next slot, and overhear, e.g. the possibility of earning money). Further, we avoided selection into treatment by not announcing treatments ex ante and randomly assigning treatments to days after most booking slots had already been filled.[13]

Upon arrival, *ExitTheRoom* staff welcomed teams of customers as usual and customers signed ETR's terms and conditions, including ETR's data privacy policy. Then, the staff explained the rules of the game. Afterwards, the teams were shown to their room and began solving the task. Teams were not informed that they were taking part in an experiment. The only difference between the treatment conditions and the control was that in the bonus conditions, the bonuses were announced as a special offer to reward particularly successful teams, while in the reference point treatment, the finishing time of 45 minutes was mentioned saliently before the team started working on the task.

2.3.2 Framed Field Experiment (Student Sample)

For the framed field experiment, we invited student participants from the social sciences laboratory at the University of Munich (MELESSA). Between March and June 2016, and January and May 2017, a total of 804 participants (268 groups) took part in the experiment. To avoid selection into the sample based on interest in the task, we recruited these participants using a neutrally framed invitation text that did not explicitly state what activity participants could expect. The invitation email informed potential participants that the experiment consisted of two parts, of which only the first part would be conducted on the premises of MELESSA whereas the second part would take place outside of the laboratory (without mentioning the escape game). They were further informed that their earnings from the first part would depend on the decisions they made and that the second part would include an activity with a participation fee that would be covered by the experimenters (as part of participants' compensation for taking part in the experiment).[14]

Upon arrival at the laboratory, the participants were informed about their upcoming participation in an escape game. The participants had the option to opt out of the experiment, but no one did so. In the first part of the experiment, i.e. on the premises of

[13] All slots in November and December 2015 were fully booked before treatment assignment: according to the provider, fewer than five percent of their bookings are made on the day of an event after the first time slot has ended.

[14] Section A.1 in the Appendix provides a translation of the text of the invitation.

MELESSA, we elicited the same control variables as for the customer sample (age, gender, and potential experience with escape games). In addition, the participants took part in three short experimental tasks and answered several surveys. As the main focus of this paper is to analyze the robustness of the incentive effects across the two samples, we relegate the discussion of the results from these additional tasks to another paper.[15] After completion of the laboratory part, the experimenters guided the participants to the facilities of ETR which are located a ten-minute walk (0.4 miles / 650 meters) away from the laboratory. At ETR, each participant was randomly allocated to a team of three members, received the same explanations from the ETR staff that were given in the field experiment, and, depending on the treatment, was informed about the possibility of earning a bonus. For the student sample, we randomized the treatments on the session level (stratifying on rooms), as student teams in different sessions on a given day could not talk to each other at the facilities of ETR. During the performance of the task, the same information about the team performance as in the field experiment was collected. On completion of the task, the participants answered questions about the team's behavior, organization, and their perception of the task individually, on separate tablet computers. At the end, we paid the earnings individually in cash. In addition to the participation fee for ETR, which we covered (given the regular price, this corresponds to roughly €25 per person), participants earned on average €7.53, with payments ranging from €3.50 to €87.[16]

3 Results

We organize the presentation of our findings as follows. We begin our analysis by establishing the internal validity of our experimental approach. We show that the student participants perceive the task at *ExitTheRoom* as non-routine and analytical, i.e. involving

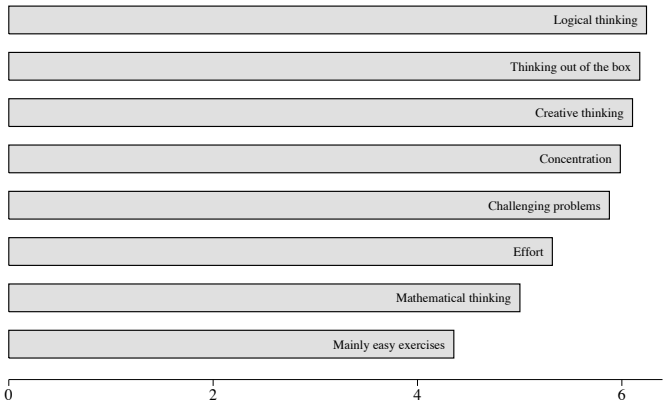
[15] These tasks included an elicitation of the willingness-to-pay for a voucher of *ExitTheRoom*, an experimental measure of loss aversion (based on Gächter et al. (2007)) and a word creation task (developed by Eckartz et al. (2012)). The participants also answered questionnaires regarding creativity (Gough, 1979), competitiveness (Helmreich and Spence, 1978), status (Mujcic and Frijters, 2013), a big five inventory (Gosling et al., 2003), risk preferences (Dohmen et al., 2011) and standard demographics. On average, the subjects spent roughly 30 minutes to complete the experimental tasks and questionnaires.

[16] In one of the laboratory tasks, the student participants further had the chance to win a voucher for ETR worth roughly €100. Twenty-six participants actually won such a voucher, implying an average additional earnings from this task of roughly €3.23. Adding up all these earnings assuming market prices as valuations, the participants on average earned an equivalent of €35.76 for an experiment lasting two hours.

more cognitive effort and creative thinking than easy, routine exercises. Then, we analyze our main research question, whether bonuses improve team performance. As our findings are affirmative, we explore next the channels through which bonus incentives operate. We disentangle which elements of the bonus (framing, monetary reward, reference point) are most relevant for bringing about the performance effect and investigate whether the observed effects of bonuses on performance are robust. We study whether the effects of bonuses on the teams that self-selected into the task differ from those on the teams that we confronted with the task, and whether the bonuses affect team organization. Finally, we highlight how bonus incentives affect a team's willingness to explore new approaches, and evaluate whether incentives affect this exploratory behavior differently for teams in the natural versus the framed field experiment.

3.1 Task Perception and Randomization

We have previously argued that real-life escape games offer the opportunity to study a class of tasks that is highly relevant to modern workplaces, as teams face a non-routine, analytical, and interactive challenge that requires thinking outside the box and logical thinking rather than easy repetitive chores. In order to not interfere with the standard procedures at *ExitTheRoom*, we could not run extensive surveys and, e.g., ask regular customers about their perception of the task. However, we asked the student participants from the framed field experiment ($N = 804$) to what extent they agree that the team task exhibits various characteristics (using a seven-point Likert scale). Figure 2 shows the mean answers of our participants. Participants strongly agreed that the task involves logical thinking, thinking outside the box, and creative thinking, in particular as compared to mathematical thinking and easy exercises (signed-rank tests reject that the ratings have the same underlying distribution, all p-values < 0.01 except for *Thinking outside the box* vs. *Logical thinking*, $p = 0.16$ and *Thinking out of the box* vs. *Creative thinking* $p = 0.02$).



The figure shows mean answers of $N = 804$ student participants to eight questions concerning attributes of the task. Answers were given on a 7-point Likert scale.

Figure 2: Task perception

Table 1: Sample size and characteristics

	Control (n=238)	*Bonus45 (pooled) (n=249)*
Share males	0.52 (0.29) [0,1]	0.51 (0.29) [0,1]
Group size	4.53 (1.18) [2,7]	4.71 (1.05) [2,8]
Experience	0.48 (0.50) [0,1]	0.48 (0.50) [0,1]
Private	0.69 (0.46) [0,1]	0.63 (0.48) [0,1]
English speaking	0.12 (0.32) [0,1]	0.08 (0.28) [0,1]
Age category $\in \{$18-25;26-35;36-50;51+$\}$	$\{$0.29;0.45;0.21;0.05$\}$	$\{$0.18;0.42;0.33;0.07$\}^{***}$

All variables except age category refer to means on the group level. Experience refers to teams that have at least one member who experienced an escape game before. Private refers to whether a team is composed of private members (1) or whether the team belongs to a team building event (0). Standard deviations and minimum and maximum values in parentheses; (std.err.)[min, max]. Age category displays fractions of participants in the respective age category. Stars indicate significant differences to *Control* (using χ^2 tests (for frequencies) and Mann–Whitney tests (for distributions), with $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

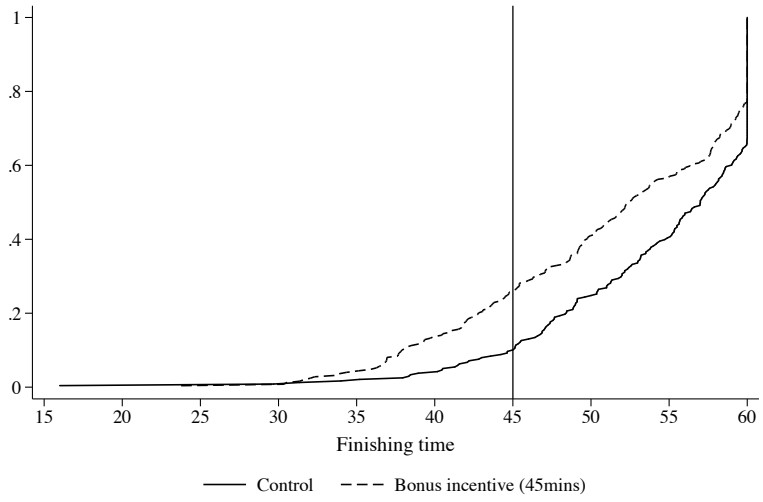
Table 1 provides an overview of the properties of the sample in the main treatments of the natural field experiment with *ETR* customers. The table highlights that our randomization was successful, based on observables such as the share of males, group size, experience, whether teams were taking part in a private or company event, and whether the team was German-speaking.

The only characteristic which differs significantly across treatments is the distribution of participants over the age categories guessed by our research assistants (χ^2 test, p-value < 0.01). We therefore provide results from both the regression specifications

without controls and the regression specifications in which we control for the estimated age ranges (and other observables).

3.2 Bonus Incentives and Team Performance

We now turn to our primary research question, whether providing bonus incentives improves team performance. As mentioned earlier, our objective outcome measure of performance is whether teams manage to solve the task within 45 minutes and more generally how much time teams need to solve the task. Figure 3 shows the cumulative distribution of finishing times with and without bonus incentives in the field experiment. The vertical line marks the time limit for the bonus. The figure indicates that bonus incentives induce teams to complete the task faster and that the positive effect is not only prevalent around the bonus threshold but over a large part of the support of the distribution.



The figure shows the cumulative distributions of finishing times with and without bonus incentives. The vertical line marks the time limit for the bonus.

Figure 3: Finishing times in *Bonus45* and *Control* in the field experiment

In *Control*, only 10 percent of the teams manage to finish the task within 45 minutes whereas in the bonus treatments more than twice as many teams (26.1 percent) do so (χ^2 test, p-value < 0.01). The remaining time upon solving also differs significantly between *Bonus45* and *Control* (p-value < 0.01, Mann–Whitney test). In the bonus treatment, teams are on average about three minutes faster than in *Control*. The positive

effect of bonuses on performance is also reflected in the fraction of teams finishing the task within 60 minutes. With bonuses, 77 percent of the teams finish the task before the 60 minutes expire, whereas in *Control* this fraction amounts to only 67 percent (χ^2 test, p-value $= 0.01$, see also Table 4).

In addition to our non-parametric tests, we provide regression analyses which allow us to control for observable team characteristics (gender composition of the team, team size, experience with escape games, private vs. team building, English-speaking, and the estimated age of team members). Table 2 presents the results from a series of probit regressions that estimate the probability of solving the task within 45 minutes. To provide against heteroskedasticity, we employ Huber–White standard errors throughout. Column (1) includes only a dummy variable for the bonus treatments *Bonus45*. Bonus incentives are estimated to increase the probability of solving the task in less than 45 minutes by 16.5 percentage points. In Column (2), we add the observable characteristics mentioned above (see also Table 1). Here, and in the following analysis, group size and experience with escape games have a positive effect on performance whereas English speaking groups perform slightly worse.[17] In Column (3) we add fixed effects for the ETR staff members on duty and in Column (4) we add week fixed effects. Across all specifications, the coefficients of the bonus treatments are positive and highly significant. Paying bonuses to teams solving a non-routine task strongly enhances their performance. We also estimate the effects of bonuses on the time remaining upon solving the task, which largely confirms both the results from the non-parametric tests on the remaining time as well as the results from the Probit models in Table 2, although the results are not statistically significant in all specifications (see Table A.2 in Appendix A.3.2).

We can look in more detail at the effectiveness of incentives depending on time elapsed since the beginning of the task. Since the incentive only rewards completing the task in the first 45 minutes, it should theoretically lose its effect in the last 15 minutes of the task. In addition, if incentives crowd out out intrinsic motivation to solve the task, we should see a decrease in performance after 45 minutes compared to *Control*. To test this hypothesis, we run a Cox proportional hazard model, where we define the hazard as completing the task. If our prior was true, we should observe the treatment to have a

[17] See also Table A.4 in the Appendix. Note further that the treatment effect does not strongly interact with the observable team characteristics. Only the interaction of incentives and experience (model (4) in A.4) turns out to be significantly positive at the ten percent level.

strong effect on the hazard in the first 45 minutes, no or even a negative effect in the last 15 minutes, conditional on covariates.

Table 2: Probit regressions (ME) on solved in less than 45 minutes

	Probit (ME): Solved in less than 45 minutes				
	(1)	(2)	(3)	(4)	(5)
Bonus45 (pooled)	0.165***	0.164***	0.188***	0.151***	
	(0.033)	(0.034)	(0.037)	(0.056)	
Gain45					0.125*
					(0.064)
Loss45					0.174***
					(0.061)
Fraction of control teams solving the task in less than 45 min	0.10	0.10	0.10	0.10	0.10
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Observations	487	487	487	487	487

The table displays average marginal effects from Probit regressions of whether a team solved the game within 45 minutes on our treatment indicator (with *Control* as base category). Control variables added from column (2) onwards include team size, share of males in a team, a dummy whether someone in the team has been to an escape game before, dummies for median age category of the team, a dummy whether all group members speak German and a dummy for private teams (opposed to company team building events). Staff fixed effects control for the employees of *ExitTheRoom* present onsite and week fixed effects for week of data collection. All models include the full sample, including weeks that perfectly predict failure to receive the bonus (Table A.1 in section A.3 of the Appendix reports regressions from a sample excluding weeks without variation in the outcome variable). Robust standard errors reported in parentheses, with $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

Table 3 shows the hazard ratios using our usual set of controls and employing robust standard errors. Columns (1) through (3) estimate the effect on the hazard rate for the first 45 minutes and columns (4) through (6) on the last 15 minutes. In columns (1) and (4) we present the baseline effect of the treatment without any covariates. These are added in columns (2) and (5) respectively. Columns (3) and (6) also include week and staff fixed effects. The treatment clearly increases the hazard rate of completing the task in the first 45 minutes. All coefficients are significantly different from 1 and large in magnitude. Adding controls and fixed effects doesn't change the estimates by much, and the p-values of the proportional hazard assumption test do not indicate any reason to doubt our specification. In the last 15 minutes (columns (4) to (6)), however, the effect has almost completely vanished. The coefficient on our treatment switches from far above one to around one, and is marginally significant in only one out of three specifications. Again, the proportional hazard assumption cannot be rejected. Thus our data reflects two important aspects: First, the treatment indeed increases the likelihood of completing the

task in the first 45 minutes, but much less so in the last 15 minutes. Second, incentives are unlikely to crowd out intrinsic motivation in our setting. We conclude:

Result 1 *Bonus incentives increase team performance in the non-routine task.*

Table 3: Influence of treatment on hazard rates

	Cox Proportional Hazard Model: Finishing the Game					
	First 45 minutes (1)-(3)			Last 15 minutes (4)-(6)		
	(1)	(2)	(3)	(4)	(5)	(6)
Bonus45 (pooled)	2.853***	2.947***	2.914***	1.178	1.250*	0.841
	(0.680)	(0.718)	(1.371)	(0.145)	(0.165)	(0.214)
p-value for prop. haz. assumption	0.743	0.479	0.447	0.845	0.540	0.631
Control Variables	No	Yes	Yes	No	Yes	Yes
Staff Fixed Effects	No	No	Yes	No	No	Yes
Week Fixed Effects	No	No	Yes	No	No	Yes
Observations	487	487	487	487	487	487

Hazard ratios from a Cox proportional hazard regression of time elapsed until a team has completed the task on our treatment indicator *Bonus45*. Control variables, staff and week fixed effects as in Table 2. Robust standard errors reported in parentheses, with $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$. Significant coefficients imply that the null hypothesis of equal hazards (i.e. ratio = 1) can be rejected. The proportional hazard assumption is tested against the null that the relative hazard between the two treatment groups is constant over time.

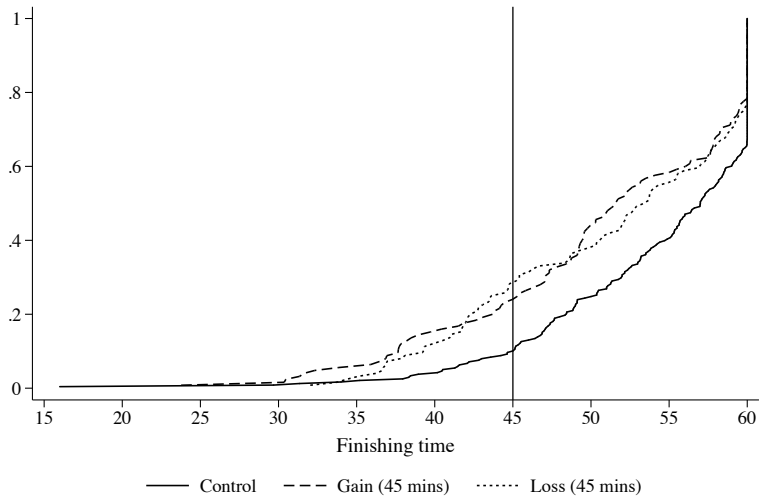
3.3 Elements of Bonus Incentives: Framing, Rewards and Reference Performance

3.3.1 Framing of Bonus Incentives

As explained in the section on the experimental design, for roughly one-half of the teams in *Bonus45* we framed the bonus incentives as gains, while the other half faced a loss frame. Figure 4 shows the cumulative distributions of finishing times separately for both frames. We find that the framing of the bonus is of minor importance for team performance. A Mann–Whitney test fails to reject the null hypothesis that the finishing times for the two framings come from the same underlying distribution (p-value $= 0.70$). Also, the fractions of teams solving the task within 45 minutes does not differ significantly (in *Gain45*, 24 percent of teams finish within 45 minutes, in *Loss45* 28 percent of teams do so, χ^2-test, p-value $= 0.45$). Further, the fraction of teams solving the task in 60 minutes (78 percent in *Gain45* and 77 percent in *Loss45*) does not differ significantly (χ^2-test, p-value $= 0.85$) and no statistically significant differences are observed for the remaining times

across frames: In *Gain45*, teams have on average 36 seconds more left than in *Loss45*, and the successful teams in *Gain45* have on average 37 seconds more left than in *Loss45* (Mann–Whitney test, p-value $= 0.71$). Table 4 summarizes these different performance measures. In addition to the non-parametric analyses we report results from a regression of the probability of solving the task within 45 minutes on a separate dummy for each framing of the bonus and our control variables in Column (5) of Table 2. Incentives significantly increase the probability of solving the task within 45 minutes under both frames (as compared to the control condition) but a post-estimation Wald test shows that there is no statistically significant additional impact from framing the bonus as a loss (p-value$= 0.38$). We summarize these findings in Result 2.

Result 2 *Framing the bonus as a loss has no significant additional advantage over framing the bonus as a gain.*



The figure shows the cumulative distribution of finishing times with bonus incentives framed as either gains, losses, or without bonuses. The vertical line marks the time limit for the bonus.

Figure 4: Finishing times in bonus treatments (disaggregated) and *Control* in the field experiment

3.3.2 Reference Points vs. Monetary Rewards

To understand whether bonus incentives work due to the monetary reward or due to the fact that the bonus also created a salient reference point at the 45-minute mark, we con-

Table 4: Task performance with and without bonus incentives

	Control	*Bonus45 (pooled)*	*Gain45*	*Loss45*
fraction of teams solving task in 45 mins	0.10	0.26^{***}	0.24^{***}	0.28^{***}
fraction of teams solving task in 60 mins	0.67	0.77^{**}	0.78^{**}	0.77^{*}
mean remaining time (in sec)	345	530^{***}	548^{***}	512^{***}
mean remaining time (in sec) if solved	515	688^{***}	707^{***}	669^{***}

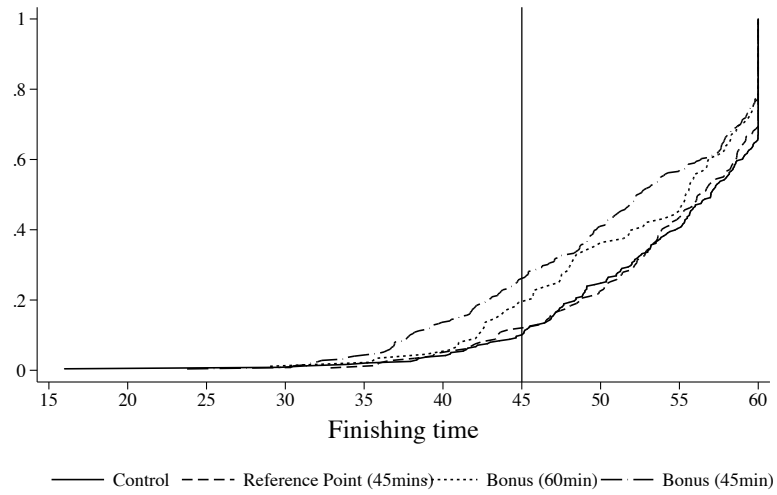
This table summarizes key variables and their differences across our three treatments *Control*, *Gain45*, and *Loss45* and the pooled bonus incentive treatments. Stars indicate significant differences from *Control* (using Fisher's exact test for frequencies and Mann–Whitney tests for distributions), with $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

ducted two additional treatments. In *Reference Point* we introduce the 45-minute threshold as a salient reference point but do not pay a reward. In *Bonus60* we pay a bonus (again framed as a gain or a loss) for solving the task in 60 minutes.[18] Figure 5 shows the cumulative distribution of finishing times in *Control*, *Reference Point*, *Bonus60* and *Bonus45* and indicates that monetary rewards reduce the amount of time teams need to finish the task (*Bonus60* vs. *Control*, Mann–Whitney test, p-value $= 0.05$; *Bonus45* vs. *Control*, Mann–Whitney test, p-value < 0.01, with *Bonus45* vs. *Bonus60*, Mann–Whitney test, p-value $= 0.24$), whereas the cumulative distribution of remaining times in *Reference Point* almost perfectly overlaps with the cumulative distribution function in *Control* (Mann–Whitney test, p-value $= 0.78$). Hence, this is strong evidence that it is not the provision of a salient reference performance, but rather the reward component of the bonus incentives which generates the performance increase.

Lastly, we provide a regression analysis for the full sample in Table 5. We regress the probability of finishing within 45 minutes on the three treatment indicators *Reference Point*, *Bonus60* and *Bonus45*. Column (1) includes only the treatment dummies. In Column (2), we add our set of control variables. In Column (3) we add staff fixed effects and in Column (4) we add week fixed effects. The regressions show that monetary incentives significantly increase the probability of finishing within 45 minutes, whereas the reference treatment does not.[19] It also becomes apparent that this finding is robust to the addition of covariates and fixed effects. Moreover, a post-estimation Wald test rejects the equality of coefficients of *Bonus60* and *Reference Point* in all specifications controlling for covariates (models (2) to (4), p-values$<$0.1) but fails to reject equality of coefficients

[18] We do not differentiate between the gain and the loss frame of *Bonus60* in the following. As for *Bonus45*, no difference between the frames emerged.

[19] Table A.3 in Appendix A.3 confirms these findings for remaining time as dependent variable.



The figure shows the cumulative distribution of finishing times of all bonus treatments (45 minutes and 60 minutes pooled each), *Reference Point* and *Control*. The vertical line marks the time limit for the *Bonus45* condition.

Figure 5: Finishing times for all treatments in the field experiment

at conventional levels of statistical significance (p-value=0.11) for model (1), which includes no covariates. Similarly, the coefficient of *Bonus45* is significantly larger than the coefficient of *Reference Point* (at the 1 percent level) except for the specification in column (4) (p-value=0.14). Equality of coefficients of *Bonus60* and *Bonus45* can never be rejected. We summarize this finding in Result 3:

Result 3 *Bonuses increase performance due to the monetary reward they provide. Introducing a salient reference performance (indicating extraordinary performance) is not sufficient to induce a performance shift.*

Table 5: Probit regressions (ME) on solved in less than 45 minutes (all treatments)

	Probit (ME): Solved in less than 45 minutes			
	(1)	(2)	(3)	(4)
Bonus45 (pooled)	0.160***	0.157***	0.164***	0.108**
	(0.033)	(0.033)	(0.035)	(0.047)
Bonus60 (pooled)	0.105**	0.102**	0.105**	0.127**
	(0.046)	(0.044)	(0.046)	(0.059)
Reference Point	0.025	0.023	0.011	0.020
	(0.042)	(0.041)	(0.045)	(0.052)
Fraction of control teams solving the task in less than 45 min	0.10	0.10	0.10	0.10
Control Variables	No	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes
Week Fixed Effects	No	No	No	Yes
Observations	722	722	722	722

The table shows average marginal effects from Probit regressions of whether a team solved the task within 45 minutes on our treatment indicators *Bonus45*, *Bonus60* and *Reference Point* with *Control* being the base category. Control variables, staff and week fixed effects as in Table 2. Robust standard errors reported in parentheses, and $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

3.4 Robustness of the Bonus Incentive Effect: Results from the Framed Field Experiment

We have shown that bonus incentives increase performance in our non-routine team task in a sample of self-selected and motivated teams of *ETR* customers. To test whether the performance enhancing effect of bonus incentives in non-routine analytical team tasks is also present in demographics other than the self-selected ETR customer sample, we repeated our main treatments in a student sample. Student participants may react differently to bonus incentives than the teams from our natural field experiment for several reasons. Most importantly, the process by which the sample is drawn is different across the two experiments. While regular teams of *ExitTheRoom* customers self-select into the task and are likely to be intrinsically motivated to perform well (as they pay for it), student teams from the laboratory subject pool are confronted by us with the task, do not pay for it, and hence are less likely to be intrinsically motivated to solve the task. Teams in the field experiment are also formed endogenously and vary in size, whereas we randomly as-

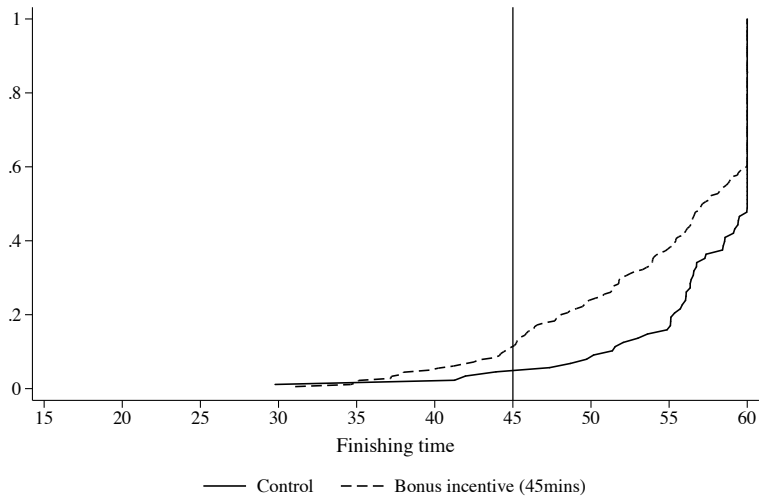
sign students to teams of three participants. Finally, our student participants differ along several observable dimensions, such as age, gender and experience with the task.[20]

In all, we randomized 268 teams of three students into the treatments *Control* (88), *Gain45* (90) and *Loss45* (90). Despite the assignment to the treatment being random and balanced across weeks, there are on average fewer males in *Gain45* (0.39) than in *Control* (0.46) (Mann–Whitney test, *Gain45* vs. *Control*, p-value $= 0.08$) or *Loss45* (0.47) (Mann–Whitney test, *Loss45* vs. *Control* p-value $= 0.10$, *Loss45* vs. *Gain45*, p-value $= 0.97$), and the share of teams with at least one team member with experience in escape games is higher in *Loss45* (0.42) than in *Gain45* (0.29) (χ^2- test, p-value $= 0.06$). Age does not significantly differ by treatment (Mann–Whitney test, *Gain45* vs. *Control* p-value $= 0.47$, *Loss45* vs. *Control*, p-value $= 0.92$ and *Loss45* vs. *Control*, p-value $= 0.38$). Although the differences between treatments are not very pronounced, we will nevertheless control for these differences in our regression analyses.

Analogously to the analysis in the customer sample, we study treatment effects on team performance by analyzing the fraction of the teams solving the task in 45 minutes, and 60 minutes respectively, as well as the remaining times of teams in general and among successful teams. Figure 6 shows the performance of teams in the framed field experiment and is the student sample analogue to Figure 3. While student teams perform worse on average than the ETR customer teams, the bonus incentives turn out to be similarly effective for the student teams.

Again, the fraction of teams finishing within 45 minutes is more than twice as high when teams face bonus incentives. In the incentive treatments, 11 percent of teams manage to solve the task within 45 minutes whereas only 5 percent do so in *Control* (χ^2-test, p-value $= 0.08$). The fraction of teams finishing the task within 60 minutes is also significantly larger under bonus incentives. With bonuses, 60 percent of the teams finish the task before the 60 minutes expire whereas in *Control* this fraction amounts to 48 percent (χ^2-test, p-value $= 0.06$). Further, with bonus incentives teams are on average about three minutes faster than in *Control*, and Mann–Whitney tests reject that finishing times in the control condition come from the same underlying distribution as finishing times under bonus incentives (Mann–Whitney test, p-values < 0.01). Table 6 summarizes these findings.

[20]The students are on average younger (23.03), slightly less likely to be male (44 percent) and less experienced in escape games (36 percent of the student teams had at least one member with experience in escape games).



The figure shows the cumulative distributions of finishing times. The vertical line at 45 minutes marks the time limit for the bonus.

Figure 6: Finishing times across treatments in the framed field experiment (student sample)

In addition to the non-parametric tests, we run regressions analogously to the analyses for the customer sample. As before, we control for the share of males in a team, average age and experience with escape games.[21] Table 7 reports the results from Probit regressions on the probability of solving the task within 45 minutes. Column (1) only uses the treatment dummy and shows that bonus incentives significantly increase the probability of solving the task in 45 minutes. The positive effects of the bonus incentives are robust to controlling for background characteristics (Column (2)), for staff fixed effects (Column (3)), and week fixed effects (Column (4)). Overall, the Probit regression results reinforce our non-parametric findings. Offering bonuses increases team performance. Running a regression separately for gain and loss frames yields qualitatively very similar results (Column (5)), as the coefficients for *Loss45* and *Gain45* are again both positive. However, only the coefficient for the gain frame turns out to be statistically significant. A post-estimation Wald test cannot reject equivalence for the coefficients of *Gain45* and *Loss45* at the ten percent level. Also for the student sample, the positive effect of bonus

[21] In contrast to the ETR customer sample all teams speak German and consist of three team members. Hence, we do not need to control for language or group size.

incentives is reflected qualitatively in the analyses of the time remaining (see Table A.6 in Appendix A.4).

Table 6: Task performance with and without bonus incentives (student sample)

	Control	*Bonus45 (pooled)*	*Gain45*	*Loss45*
fraction of teams solving task in 45 mins	0.05	0.11^{*}	0.13^{**}	0.09
fraction of teams solving task in 60 mins	0.48	0.60^{*}	0.54	0.66^{**}
mean remaining time (in sec)	169.90	327.97^{***}	321.28^{*}	334.67^{***}
mean remaining time (in sec) if solved	355.98	546.62^{***}	590.10^{**}	510.50^{***}

This table summarizes key variables and their differences across our three treatments *Control*, *Gain45* and *Loss45*, as well as the combined *Bonus45 (pooled)*. Stars indicate significant differences from *Control* (using χ^2 test for frequencies and Mann–Whitney tests for distributions), with $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$. P-values of non-parametric comparisons between *Gain45* and *Loss45* exceed 0.10 for all four performance measures.

Table 7: Probit regressions (ME) on solved in less than 45 minutes (student sample)

	Probit (ME): Solved in less than 45 minutes				
	(1)	(2)	(3)	(4)	(5)
Bonus45 (pooled)	0.075^{*}	0.073^{*}	0.075^{*}	0.079^{**}	
	(0.042)	(0.042)	(0.041)	(0.039)	
Gain45					0.101^{**}
					(0.043)
Loss45					0.051
					(0.041)
Fraction of control teams solving the task in less than 45 min	0.045	0.045	0.045	0.045	0.045
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Observations	268	268	268	268	268

The table shows average marginal effects from Probit regressions of whether a team solved the game within 45 minutes on our treatment indicator (with *Control* as base category). Control variables added from column (2) onwards include share of males in a team, a dummy whether someone in the team has been to an escape game before and average age of the team. Staff fixed effects control for the employees of *ExitTheRoom* present onsite and week fixed effects control for week of data collection. All models include the full sample, including weeks that perfectly predict failure to receive the bonus (Table A.5 in section A.3 of the Appendix reports regressions from a sample excluding weeks without variation in the outcome variable). Robust standard errors reported in parentheses, with $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

3.5 Performance and Team Organization

In addition to establishing the robustness of the positive incentive effect, our student sample allows us to explore whether bonus incentives also affect team motivation and organization. We conducted two post-experimental questionnaires to analyze potential

mechanisms through which the treatment effect could operate. In Questionnaire 1, we asked our student participants to agree or disagree (on a seven-point Likert scale) with a number of statements that might capture aspects of team motivation and organization. In Questionnaire 2 (which was conducted for a subsample of 375 participants), we use an additional set of questions based on the concept of team work quality by Hoegl and Gemuenden (2001). Table 8 reports the results from Questionnaires 1 and 2.

The upper panel of Table 8 shows that incentives in general do not strongly affect agreement with the statements we provided but reveals some interesting insights, about through which channels incentives might potentially operate. First, teams appear to be notably more stressed when facing incentives than were the teams in the control (Mann–Whitney test, p-value $= 0.01$). At the same time, similar to teams in *Control*, treated teams strongly agree with the statement "I would like to participate in a similar task again" (Mann–Whitney test, p-value $= 0.88$), suggesting that incentives caused positive rather than negative stress among the team members. Second, participants in the incentive treatment are more likely to report that one team member was dominant in leading the team (Mann–Whitney test, p-value $= 0.03$), and also agree significantly more with the statement "I was dominant in leading the team" (Mann–Whitney test, p-value $= 0.05$). Additionally, we observe several differences in items relating to a more focused and directed approach within a team (although some of them fail to be statistically significant at the 10 percent level): With bonus incentives, participants tend to agree less with the statements "We wrote down all numbers we found." (Mann–Whitney test, p-value $= 0.04$), "We exchanged many ideas in the team." (Mann–Whitney test, p-value $= 0.12$) and "When we got stuck we let as many team members try as possible." (Mann–Whitney test, p-value $= 0.14$).

Table 8: Answers to post-experiment questionnaires

	Control	*Bonus45*	p-values (Mann–Whitney)
Questionnaire 1 (n=804)			
"The team was very stressed."	3.57	4.13***	0.00
"One person was dominant in leading the team."	2.60	2.86**	0.03
"We wrote down all numbers we found."	5.64	5.50**	0.04
"I was dominant in leading the team."	2.64	2.87**	0.05
"We first searched for clues before combining them."	4.58	4.39	0.11
"We exchanged many ideas in the team."	5.87	5.74	0.12
"When we got stuck we let as many team members try as possible."	5.43	5.28	0.14
"The team was very motivated."	6.14	6.26	0.22
"We communicated a lot."	5.78	5.88	0.23
"All team members exerted effort."	6.23	6.37	0.24
"Our notes were helpful in finding the solution."	5.50	5.43	0.41
"I was able to present all my ideas to the group."	5.95	5.93	0.41
"We were well coordinated in the group."	5.73	5.80	0.61
"I was too concentrated on my own part."	2.88	2.83	0.76
"We made our decisions collectively."	5.51	5.58	0.87
"I would like to perform a similar task again."	6.30	6.28	0.88
"Our individual skills complemented well."	5.65	5.68	0.89
"The mood in our team was good."	6.30	6.36	0.93
"All team members contributed equally."	5.97	6.00	0.96
Questionnaire 2 (n=375)			
"How much did you wish somebody would take the lead?"	2.67	3.32***	0.00
"How well led was the team?"	3.85	4.21**	0.04
"How much did you think about the problems?"	6.00	5.79	0.11
"How much did you follow ideas that were not promising?"	5.02	4.79	0.17
"How much team spirit evolved?"	5.54	5.80	0.17
"How much coordination was there of individual tasks and joint strategy?"	3.28	3.51	0.18
"How much exploitation was there of individual potential?"	5.14	4.94	0.22
"How much helping was there when somebody stuck?"	5.70	5.58	0.22
"How much did you search the room for solutions?"	6.31	6.22	0.51
"How much exertion of effort was there by all the members?"	5.98	5.96	0.60
"How much communication was there about procedures?"	5.30	5.35	0.88
"How much was there of accepting the help of others?"	5.80	5.85	0.89

This table reports answers to our post-experiment questionnaires from the framed field experiment by treatment (*Control* and *Bonus45*), and p-values of the differences between the treatments. The scale ranges from not at all agreeing to the statement (=1) to completely agreeing (=7) in Questionnaire 1 and from very little (=1) to very much (=7) in Questionnaire 2. Stars indicate significant differences from *Control* using Mann-Whitney tests, with $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

The results from Questionnaire 2 in the lower panel of Table 8 mirror the answers from Questionnaire 1. Teams facing incentives report more demand for leadership (Mann–Whitney test, p-value < 0.01), while they also report that teams were better led (Mann–

Whitney test, p-value$= 0.04$). Further, also in Questionnaire 2 we observe several tendencies suggesting a potentially more focused and directed approach within the teams under incentives. Teams tend to be less likely to spend a long time thinking about problems (Mann–Whitney test, p-value $= 0.11$) and tend to follow ideas that were not promising less frequently (Mann–Whitney test, p-value $= 0.17$) . Also, teams facing bonus incentives tend to be more likely to report an emergence of team spirit (Mann–Whitney test, p-value $= 0.17$) and the coordination of individual tasks and joint strategy (Mann–Whitney test, p-value $= 0.18$). Although these statistically insignificant results can serve as suggestive evidence only, we nonetheless believe that they highlight a potentially relevant channel through which bonus incentives for teams may increase performance: with an incentive, teams demand more leadership, individual team members are more likely to take the initiative and teams become more focused and better coordinated.

3.6 Bonus Incentives and the Willingness to Explore

The effectiveness of bonus incentives in the long run depends on whether monetary incentives crowd out intrinsic motivation, thereby inhibiting creativity and innovation. In fact, previous research has suggested that performance-based financial incentives may do just that, and thereby affect workers' willingness to explore in an experimentation task (see, e.g., Ederer and Manso, 2013). Our setup allows us to shed light on whether such behavioral reactions are also present in the context of non-routine analytical team tasks. We interpret the request for external help (hint taking) as a proxy for a team's unwillingness to explore on their own, and thus analyze how many out of the five possible hints teams request under the different treatment conditions, as well as whether they are more likely to take hints earlier in the presence of incentives.

Table 9 shows the number of hints taken across samples and treatments. For teams who self-selected into the task (customer sample), we do not find a statistically significant difference in the number of hints taken within 60 minutes. These teams take on average about three hints in both the bonus treatment and the control condition. In contrast, for teams confronted by us with the task (the student sample), we observe (economically and statistically) significantly more hint taking in the bonus treatments than in *Control*, suggesting that incentives reduce these student teams' willingness to explore original solutions. To capture potential heterogeneity across teams, we report the fractions of

teams requesting 0, 1, 2, 3, 4 or 5 hints for the customer sample in panel (a) and for the student sample in panel (b) of Figure 7. The figure reinforces our earlier findings: bonus incentives have, if at all, a minor effect on the number of hints taken in the customer sample. These teams' willingness to explore original solutions fails to differ statistically significantly across treatments (χ^2-test, p-value=0.114). Panel (b) of Figure 7 depicts the same histogram for the framed field experiment with student participants. It becomes apparent that teams who did not self-select into the task are much more likely to take hints when facing incentives (χ^2-test, p-value=0.029). Roughly 75 percent of these teams take four or five hints when facing incentives, as compared to 59 percent doing so in *Control*. Regression analyses on hint taking (including additional controls, see Table 10, models (1), (2), (5), and (6)) confirm these results.[22]

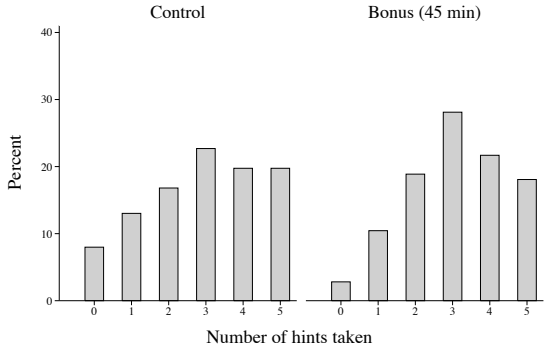
Table 9: Hints requested in the field experiment and the framed field experiment

	Control	*Bonus45 (pooled)*	*Gain45*	*Loss45*
within 60 minutes				
Field Experiment (487 groups)	2.92 (1.55)	3.10 (1.34)	3.05(1.40)	3.15(1.29)
Framed Field Experiment (268 groups)	3.74(1.04)	4.11(0.98)***	4.10(0.98)**	4.12(0.98)**
within 45 minutes				
Field Experiment (487 groups)	1.97 (1.22)	2.36 (1.15)***	2.30(1.1.19)**	2.41(1.10)***
Framed Field Experiment (268 groups)	2.33(0.93)	3.17(1.04)***	3.07(1.04)***	3.28(1.04)***

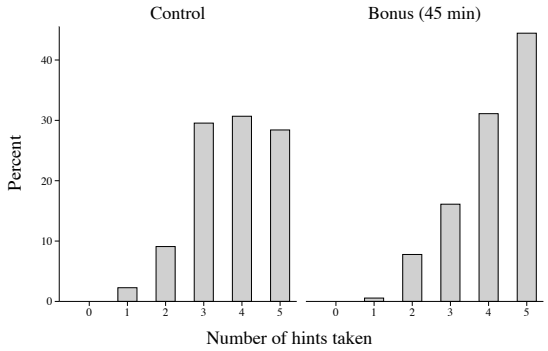
This table summarizes mean number of hints taken across treatments in the field experiment and the framed field experiment (standard deviations in parentheses). Stars indicate significant differences from *Control* (using Mann–Whitney tests), with $^* = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$. p-values of non-parametric comparisons between *Gain45* and *Loss45* are larger than 0.10 for both the field experiment and the framed field experiment.

Focusing only on hints taken within the first 45 minutes, non-parametric tests indicate significant differences across treatments for both samples, but again, the effect is much stronger for student teams who were confronted by us with the non-routine task. Regression analysis implies that these teams take on average 0.84 more hints within the first 45 minutes when facing incentives, whereas customer teams take on average only 0.39 more hints (columns (3) and (7) of Table 10). When we add additional controls and fixed effects (columns (4) and (7) of Table 10), the results for the student sample remain unchanged, whereas the positive coefficient of the incentive condition becomes statistically insignificant in the customer sample.

[22] An ordered probit regression yields qualitatively similar results, see A.10.



(a) Customer Sample (487 groups)



(b) Student Sample (268 groups)

The figure shows histograms of hints taken across samples. Panel (a) depicts the fractions of customer teams choosing 0, 1, 2, 3, 4 or 5 hints in *Control* (left graph) and *Bonus45* (right graph). Panel (b) shows the fractions for student teams.

Figure 7: Hints requested across samples and treatments

Table 10: Number of hints requested

	OLS: Number of hints requested							
	Field experiment (1)-(4)				Framed Field Experiment (5)-(8)			
	within 60 minutes		within 45 minutes		within 60 minutes		within 45 minutes	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bonus45 (pooled)	0.172	0.098	0.387***	0.186	0.372***	0.343***	0.843***	0.808***
	(0.132)	(0.221)	(0.107)	(0.192)	(0.133)	(0.131)	(0.126)	(0.125)
Constant	2.924***	4.037	1.971***	1.770	3.739***	5.449***	2.330***	4.236***
	(0.100)	(0.645)	(0.079)	(1.080)	(0.523)	(1.032)	(0.099)	(0.708)
Control Variables	No	Yes	No	Yes	No	Yes	No	Yes
Staff Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Week Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Observations	487	487	487	487	268	268	268	268

Coefficients from OLS regressions of the number of hints requested within 60 minutes or 45 minutes regressed on our treatment indicator *Bonus45*. Controls and fixed effects identical to previous tables. Robust standard errors reported in parentheses, and $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

Taken together our results are in line with the conclusion that intrinsic motivation and incentives interact in an interesting way when teams can choose whether or not to explore original and innovative solutions on their own. Customer teams who themselves chose to perform a task are presumably more intrinsically motivated to work on the task, and thus less likely to seek external help—even when facing performance incentives. In contrast, incentives strongly reduce the willingness to explore original solutions of teams that did not self-select into the task. While we are aware that the two samples differ along several other dimensions (such as exogenous versus endogenous team formation, age or educational background), it is less clear to what extent these other differences (as compared to differences in intrinsic motivation) are likely candidates to explain the differential reactions to incentives across samples. We summarize our findings in Result 4.

Result 4 *Bonus incentives reduce student teams' exploration behavior but affect exploration behavior of customer teams (if at all) to a much smaller extent.*

4 Discussion

Our results demonstrate that bonus effects have sizable effects on team performance. Importantly, these effects are present throughout all our incentive treatments, and emerge in both the natural and the framed field experiments. The performance-stimulating effect of incentives therefore seems to be ubiquitous in the non-routine analytical team task in our setting, and not simply driven by a specific choice of subjects or certain treatment parameters. The same holds for the absence of framing effects that we also observe across all treatments and samples, suggesting that framing may be specific to the environment. This is consistent with much of the literature where significant framing effects have been observed in some environments (e.g. Muralidharan and Sundararaman, 2011; Fryer et al., 2012; Hossain and List, 2012), but not in others (DellaVigna and Pope, 2017).

Further, we find that bonus incentives do not lead to strong performance decreases if teams fail to meet the time limit to receive the bonus. Instead, the proportional hazard model analysis suggests that incentives (if anything) increase the likelihood of solving the task within 60 minutes even if teams do not meet the bonus threshold of 45 minutes. Teams facing incentives (for solving the task in 45 minutes) that eventually do not obtain the bonus perform at least as well as teams not facing incentives that do not solve the task

in 45 minutes. This is particularly striking as the former are presumably more (adversely) self-selected, as the incentive effect presumably boosts some relatively good teams who would have barely missed the cutoff without incentives.

But what is driving the observed performance increase? With respect to hint-taking behavior, we have several reasons to believe that changes in hint-taking are not responsible for the observed performance effects. First, an increase in performance will mechanically make subjects request hints earlier, as they reach difficult stages earlier. Second, in our natural field experiment, overall hint-taking behavior is not significantly different across treatments. Third, when studying at what point in time teams achieve an intermediate step early in the game and how many hints teams have taken before that step, we observe significantly better performance by teams facing incentives but no significant differences in hint taking (see Table A.7 in Appendix A.5).

An alternative possible explanation for how bonuses improve performance is that incentives may enhance learning about the essentials of the production function, i.e. how combinations of different kinds of effort (e.g. searching, deliberating, combining information) map to performance. While we primarily designed our experiment with the goal of causally identifying the effect of bonus incentives, the richness of our data also allows us to shed some light on the importance of learning. We expect teams with prior experience in escape games to have acquired more knowledge on how combinations of different kinds of effort map to performance. Hence, if incentives increase performance due to learning, incentives should in particular increase the performance of inexperienced teams. However, we observe that, if at all, incentives have a stronger effect on performance of teams with prior experience (see model (4) in Table A.4), suggesting that incentives do not increase performance because of this kind of learning. While both hint-taking and learning seem unlikely to be responsible for the performance increase, we provide suggestive evidence that teams facing incentives are more likely to wish for a leader and that leaders appear to emerge endogenously when teams face incentives. This renders changes in team organization a more likely explanation for why incentives improve a team's performance.

5 Conclusion

According to Autor et al. (2003) and Autor and Price (2013), non-routine, cognitively demanding, interactive tasks are becoming more and more important in the economy. At the same time we know relatively little about how incentives affect performance in these tasks. We provide a comprehensive analysis of incentive effects in a non-routine, cognitively demanding, team task in a large scale field experiment that allows us to study the causal effect of bonus incentives on the performance and exploratory behavior of teams. Together with our collaboration partner, we were able to implement a natural field experiment with more than 700 teams and to replicate our main findings in an additional student sample of more than 250 teams. We find an economically and statistically significant positive effect of incentives on performance. Teams in both samples are more than twice as likely to solve the task in 45 minutes under the incentive condition than under the control condition, and we observe a positive performance effect not only around the bonus threshold, but for a significant part of the distribution of finishing times.

By exploiting a number of additional treatment variations in our natural field experiment, we shed more light on the drivers and moderators of the treatment effect. First, we implement the bonus incentives both in a gain and in a loss frame and find that framing team bonuses as a loss does not yield an additional performance increase as compared to framing bonuses as gains. Second, we complement the recent literature on how the provision of information about individuals' relative performance affects behavior. When providing teams with a reference point of good performance in an experimental treatment without monetary incentives, teams' finishing times do not improve compared to those in the control condition. Hence, the explicit incentives seem to be key to bringing about the positive treatment effect in our experiment. Third, we find that teams tend to be less likely to explore on their own when facing bonus incentives, but this was mostly for those teams that were mandated to perform the task. These findings extend earlier work on the (negative) relationship between incentives and the exploration of new approaches (Ederer and Manso, 2013), by highlighting a potential relationship between the consequences of incentives for exploratory behavior and the intrinsic motivation to solve a task. The fact that incentives do not always crowd out intrinsic motivation also complements recent evidence on incentive effects in meaningful routine tasks (Kosfeld et al., 2017). Finally, answers to our ex-post survey tentatively suggest that incentives may lead

to the emergence of leadership within teams in non-routine team tasks and may result in more focused approaches to work.

Our study constitutes, to the best of our knowledge, the first systematic investigation into incentive effects in non-routine analytical team tasks. The results raise interesting questions for future research. For instance, it may be promising to study explicitly how team performance in non-routine tasks changes when leadership is exogenously assigned as compared to endogenously determined. As our findings only provide an initial glimpse at the incentive effects in these kinds of tasks, systematically varying incentive structures within teams could create additional insights into the functioning of non-routine team work. Looking beyond the question of incentives, the setting of a real-life escape game may be used to study other important questions such as goal setting, non-monetary rewards and recognition, the effects of team composition, team organization, and team motivation. Studies in this setting are in principle easily replicable, many treatment variations are implementable, and large sample sizes are feasible.

References

Amabile, T. M. (1996). *Creativity in context: Update to the social psychology of creativity*. Westview Press, Boulder, Colorado.

Autor, D. H. and Handel, M. J. (2013). Putting tasks to the test: Human capital, job tasks, and wages. *Journal of Labor Economics*, 31(S1):S59–S96.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: an empirical exploration. *Quarterly Journal of Economics*, 118(4):1279–1333.

Autor, D. H. and Price, B. (2013). The changing task composition of the US labor market: An update of Autor, Levy, and Murnane (2003). Working Paper.

Azoulay, P., Graff Zivin, J. S., and Manso, G. (2011). Incentives and creativity: Evidence from the academic life sciences. *RAND Journal of Economics*, 42(3):527–554.

Bandiera, O., Barankay, I., and Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data. *Quarterly Journal of Economics*, 120(3):917–962.

Bandiera, O., Barankay, I., and Rasul, I. (2013). Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association*, 11(5):1079–1114.

Bradler, C., Neckermann, S., and Warnke, A. J. (2014). Rewards and performance: A comparison across a creative and a routine task. Working Paper.

Charness, G. and Grieco, D. (2014). Creativity and financial incentives. Working Paper.

Deci, E. L., Koestner, R., and Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6):627–668.

Delfgaauw, J. and Dur, R. (2010). Managerial talent, motivation, and self-selection into public management. *Journal of Public Economics*, 94(9):654 – 660.

Delfgaauw, J., Dur, R., Non, A., and Verbeke, W. (2015). The effects of prize spread and noise in elimination tournaments: A natural field experiment. *Journal of Labor Economics*, 33(3):521–569.

DellaVigna, S. and Pope, D. (2017). What motivates effort? evidence and expert forecasts. *Review of Economic Studies*, forthcoming.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.

Duncker, K. (1945). On problem-solving. *Psychological Monographs*, 58(5):i–113.

Eckartz, K., Kirchkamp, O., and Schunk, D. (2012). How do incentives affect creativity? Working Paper.

Ederer, F. and Manso, G. (2013). Is pay for performance detrimental to innovation? *Management Science*, 59(7):1496–1513.

Englmaier, F., Roider, A., and Sunde, U. (2017). The role of communication of performance schemes: Evidence from a field experiment. *Management Science*, 63(12):4061–4080.

Erat, S. and Gneezy, U. (2016). Incentives for creativity. *Experimental Economics*, 19(2):269–280.

Erev, I., Bornstein, G., and Galili, R. (1993). Constructive intergroup competition as a solution to the free rider problem: A field experiment. *Journal of Experimental Social Psychology*, 29(6):463–478.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, 47(1):117–132.

Friebel, G. and Giannetti, M. (2009). Fighting for talent: Risk-taking, corporate volatility and organisation change. *The Economic Journal*, 119(540):1344–1373.

Friebel, G., Heinz, M., Krüger, M., and Zubanov, N. (2017). Team incentives and performance: Evidence from a retail chain. *American Economic Review*, 107(8):2168–2203.

Fryer, R., Levitt, S., List, J., and Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. Working Paper.

Gächter, S., Johnson, E. J., and Herrmann, A. (2007). Individual-level loss aversion in riskless and risky choices. Working Paper.

Gerhart, B. and Fang, M. (2015). Pay, intrinsic motivation, extrinsic motivation, performance, and creativity in the workplace: Revisiting long-held beliefs. *Annual Review of Organizational Psychology and Organizational Behavior*, 2:489–521.

Gibbs, M., Neckermann, S., and Siemroth, C. (2017). A field experiment in motivating employee ideas. *Review of Economics and Statistics*, forthcoming.

Gittelman, M. and Kogut, B. (2003). Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns. *Management Science*, 49(4):366–382.

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528.

Gough, H. G. (1979). A creative personality scale for the adjective check list. *Journal of Personality and Social Psychology*, 37(8):1398.

Hall, B., Jaffe, A., and Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights, and methodological tools. Working Paper.

Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4):1009–1055.

Helmreich, R. L. and Spence, J. T. (1978). *The Work and Family Orientation Questionnaire: An objective instrument to assess components of achievement motivation and attitudes toward family and career*. American Psycholog. Association.

Hennessey, B. A. and Amabile, T. M. (2010). Creativity. *Annual Review of Psychology*, 61(1):569–598.

Hoegl, M. and Gemuenden, H. G. (2001). Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization Science*, 12(4):435–449.

Hossain, T. and List, J. A. (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, 58(12):2151–2167.

Jayaraman, R., Ray, D., and de Véricourt, F. (2016). Anatomy of a contract change. *The American Economic Review*, 106(2):316–358.

Kachelmaier, S. J., Reichert, B. E., and Williamson, M. G. (2008). Measuring and motivating quantity, creativity, or both. *Journal of Accounting Research*, 46(2):341–373.

Kosfeld, M., Neckermann, S., and Yang, X. (2017). The effects of financial and recognition incentives across work contexts: the role of meaning. *Economic Inquiry*, 55(1):237–247.

Laske, K. and Schroeder, M. (2016). Quantity, quality, and originality: The effects of incentives on creativity. Working Paper.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, 90(5):1346–1361.

Levitt, S. D. and Neckermann, S. (2014). What field experiments have and have not taught us about managing workers. *Oxford Review of Economic Policy*, 30(4):639–657.

McCullers, J. C. (1978). Issues in learning and motivation. In Lepper, M. R. and Greene, D., editors, *The hidden costs of reward: New perspectives on the psychology of human motivation*, pages 5–18. Psychology Press, New York.

McGraw, K. O. (1978). The detrimental effects of reward on performance: A literature review and a prediction model. In Lepper, M. R. and Green, D., editors, *The hidden costs of reward: New perspectives on the psychology of human motivation*, pages 33–60. Psychology Press, New York.

Mujcic, R. and Frijters, P. (2013). Economic choices and status: Measuring preferences for income rank. *Oxford Economic Papers*, 65(1):47–73.

Muralidharan, K. and Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, 119(1):39–77.

Nelson, R. R. and Winter, S. G. (1982). *An evolutionary theory of economic change*. Harvard University Press, Cambridge.

Pink, D. (2009). *Dan Pink on the surprising science of motivation*. TED.

Pink, D. H. (2011). *Drive: The surprising truth about what motivates us*. Penguin.

Ramm, J., Tjotta, S., and Torsvik, G. (2013). Incentives and creativity in groups. *Working Paper*.

Rosenkopf, L. and Nerkar, A. (2001). Beyond local search: Boundary-spanning, exploration and impact in the optical disc industry. *Strategic Management Journal*, 22:287–306.

Schumpeter, J. (1934). *The Theory of Economic Development*. Harvard University Press, Cambridge.

Shearer, B. (2004). Piece rates, fixed wages and incentives: Evidence from a field experiment. *Review of Economic Studies*, 71(2):513–534.

Takahashi, H., Shen, J., and Ogawa, K. (2016). An experimental examination of compensation schemes and level of effort in differentiated tasks. *Journal of Behavioral and Experimental Economics*, 61:12–19.

Weitzman, M. (1998). Recombinant growth. *Quarterly Journal of Economics*, 113(2):331–360.

A Supplementary Appendix

A.1 Text of the Invitation to Laboratory Participants

We added the following paragraph to the standard invitation to student participants in the framed field experiment:

"Notice: This experiment consists of two parts, of which only the first part will be conducted on the premises of the MELESSA laboratory. In Part 1 you will be paid for the decisions you make. Part 2 will take place outside of the laboratory. You will take part in an activity with a participation fee. Your compensation in Part 2 will be that the experimenters will pay the participation fee of the activity for you."

A.2 Treatment Form for Bonus Treatments

Bonus treatment teams had to sign the following form, indicating understanding of the treatment procedures. For teams in the loss frame, the form further included the obligation to give back the money in case the team did not qualify for the bonus. Only one member of each team signed the form and the forms differed between the customer and student sample only in the amount of the bonus mentioned (€50 for the customer sample and €30 for the student sample). Similarly, the forms of *Bonus45* and *Bonus60* only differed in the time set for receiving the bonus.

The form for *Gain45* said:

"As usual, you have one hour in total to escapre from the room. Furthermore, we have a special offer for you today: if you escape from the room within 45 minutes, you will receive €50."

The form for *Loss45* said:

"As usual, you have one hour in total to escape from the room. Furthermore, we have a special offer for you today: You now receive €50. If you do not escape from the room within 45 minutes, you will lose the €50."

A.3 Additional Analyses for the Field Experiment

A.3.1 Probability of Solving the Game in 45 Minutes (Field Experiment)

Table A.1 reports the results for the regression columns (1) to (5) from Table 2 excluding those weeks where we do not observe variation in the outcome variable; it confirms our previous findings.

Table A.1: Probit regressions: Excluding weeks with no variation in the outcome variable

Probit: Solved in less than 45 minutes	(1)	(2)	(3)	(4)	(5)
Bonus45 (pooled)	0.150***	0.151***	0.183***	0.163**	
	(0.037)	(0.036)	(0.040)	(0.060)	
Gain45					0.134**
					(0.070)
Loss45					0.188***
					(0.066)
Fraction of control teams solving the task in less than 45 min	0.11	0.11	0.11	0.11	0.11
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Observations	451	451	451	451	451

The table reports average marginal effects from Probit regressions of whether a team solved the game within 45 minutes on our treatment indicator (with *Control* as base category). Control variables, staff and week fixed effects as in Table 2. All models exclude weeks that perfectly predict failure to receive the bonus. Robust standard errors reported in parentheses, with $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

A.3.2 Regression Analysis for Remaining Time as Dependent Variable (Field Experiment)

We also estimate the effects of bonuses on the remaining time in seconds. Because our outcome measure is strongly right skewed and contains many zeroes (as there is no time left for those not finishing the task at all), we estimate a GLM regression with a log link, again employing heteroskedasticity-robust standard errors (Table A.2). Column (1) starts out with our baseline specification which includes a dummy for the incentive treatments (pooled) only. Bonus incentives significantly increase performance (measured by the remaining time). Analogously to our analysis in Table 2, we add the set of observable controls in Column (2). In Column (3) we add staff fixed effects. In Column (4) we present the results from an estimation that also includes week fixed effects. In this model the

coefficient is still positive but fails to be statistically significant at conventional levels (p-value=0.14). Finally, in Column (5) we include two treatment dummies to test whether gain or loss frames affect performance differently. Both coefficients are of similar size but fail to be statistically significant at conventional levels (for *Gain45*, p-value=0.19 and for *Loss45*,(p-value=0.17). Further, we cannot reject the equality of the coefficients for the *Loss45* and *Gain45* treatments (Wald test, p-value=0.99).

Table A.2: GLM regressions: Bonus incentives and remaining time

	GLM: Remaining time in seconds				
	(1)	(2)	(3)	(4)	(5)
Bonus45 (pooled)	0.432***	0.447***	0.406***	0.257	
	(0.100)	(0.098)	(0.109)	(0.174)	
Gain45					0.259
					(0.199)
Loss45					0.256
					(0.188)
Constant	5.842***	4.041***	4.251***	3.803***	3.803***
	(0.079)	(0.355)	(0.404)	(0.482)	(0.481)
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Observations	487	487	487	487	487

Coefficients from a generalized linear model regression with a log link of the remaining time on our treatment indicators (with *Control* as base category). Control variables, staff and week fixed effects as in Table 2. Robust standard errors reported in parentheses, and $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

Analogously to the Probit regressions reported in Table 5, we also run GLM specifications with the remaining time as the dependent variable (Table A.3) for the full set of treatments; this confirms our findings that incentives that include rewards increase performance whereas only mentioning the reference performance does not.

Table A.3: GLM regressions: Bonus incentives, references points and remaining time

GLM: Remaining time	(1)	(2)	(3)	(4)
Bonus (pooled, 45 min)	0.432***	0.436***	0.376***	0.244
	(0.100)	(0.097)	(0.106)	(0.150)
Bonus (pooled, 60 min)	0.233*	0.267**	0.392***	0.449**
	(0.135)	(0.120)	(0.127)	(0.185)
Reference Point (45 min)	0.002	-0.001	0.102	0.131
	(0.123)	(0.118)	(0.128)	(0.149)
Constant	5.842***	4.044***	4.225***	3.713***
	(0.079)	(0.296)	(0.342)	(0.417)
Control Variables	No	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes
Week Fixed Effects	No	No	No	Yes
Observations	722	722	722	722

Coefficients from a generalized linear model regression with a log link of the remaining time on our treatment indicators (with *Control* being the base category). Control variables, staff and week fixed effects as in Table 2. Robust standard errors reported in parentheses, and $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

A.3.3 Bonus incentives and team characteristics

Table A.4 shows the results from linear probability models estimating a dummy for whether teams solve the task within 45 minutes. Model (1) includes no interactions and uses the same variables and fixed effects as model (4) in Table 2. The effect of bonus incentives is of a similar magnitude as the average marginal effect in the Probit specification. In models (2) to (6) we add interactions with observable team characteristics. The findings from these models suggest that the treatment effect does not strongly interact with the observable team characteristics. Only the interaction of incentives and experience model (4) turns out to be significant (at the ten percent level) and positive, while at the same time the treatment dummy is still statistically significant and of large magnitude. Hence, the positive incentive effect is robust and slightly larger for teams with experience.

Table A.4: Linear Probability model: Bonus incentives and the probability of solving the task in 45 minutes or less

	OLS: Finishing in less than 45 minutes					
	(1)	(2)	(3)	(4)	(5)	(6)
Bonus45 (pooled)	0.172***	0.200**	0.023	0.120*	0.130*	0.169***
	(0.061)	(0.086)	(0.162)	(0.064)	(0.076)	(0.064)
Share males	0.102*	0.130**	0.102*	0.100*	0.105*	0.103*
	(0.058)	(0.063)	(0.058)	(0.058)	(0.059)	(0.058)
Group size	0.056***	0.056***	0.042**	0.057***	0.055***	0.056***
	(0.016)	(0.016)	(0.017)	(0.017)	(0.017)	(0.016)
Experience	0.125***	0.126***	0.126***	0.058	0.124***	0.125***
	(0.036)	(0.036)	(0.036)	(0.040)	(0.036)	(0.036)
Private	0.040	0.039	0.039	0.036	-0.001	0.039
	(0.045)	(0.045)	(0.045)	(0.045)	(0.052)	(0.045)
English speaking	-0.115**	-0.117**	-0.113*	-0.114**	-0.117**	-0.129***
	(0.058)	(0.058)	(0.058)	(0.058)	(0.057)	(0.045)
Bonus45 (pooled) ...						
... × Share males		-0.055				
		(0.114)				
... × Group size			0.031			
			(0.032)			
... × Experience				0.132*		
				(0.072)		
... × Private					0.077	
					(0.079)	
... × English speaking						0.027
						(0.112)
Constant	-0.177	-0.192	-0.109	-0.179	-0.163	-0.172
	(0.151)	(0.151)	(0.160)	(0.151)	(0.149)	(0.152)
R-squared	0.155	0.156	0.157	0.162	0.157	0.156
Staff Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	487	487	487	487	487	487

Coefficients from a linear probability model . Dependent variable: Dummy for finishing within 45 minutes. All models include staff and week fixed effects as in Table 7. Robust standard errors reported in parentheses, and $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

A.4 Additional Analyses for the Framed Field Experiment

A.4.1 Probability of Solving the Game in 45 Minutes (Framed Field Experiment)

Table A.5 reports the results for the regression columns (1) to (5) from Table 7 excluding those weeks where we do not observe variation in the outcome variable; it confirms our previous findings.

Table A.5: Probit regressions (ME): Excluding weeks with no variation in the outcome variable

Probit: Solved in less than 45 minutes	(1)	(2)	(3)	(4)	(5)
Bonus45 (pooled)	0.107*	0.097*	0.104*	0.111**	
	(0.057)	(0.056)	(0.055)	(0.054)	
Gain45					0.142**
					(0.061)
Loss45					0.072***
					(0.058)
Fraction of control teams solving the task in less than 45 min	0.06	0.06	0.06	0.06	0.06
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Observations	191	191	191	191	191

Table reports average average marginal effects from Probit regressions of whether a team solved the game within 45 minutes on our treatment indicator (with *Control* as base category). Control variables, staff and week fixed effects as in Table 7. All models exclude weeks that perfectly predict failure to receive the bonus. Robust standard errors reported in parentheses, with $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

A.4.2 Regression Analysis for Remaining Time as Dependent Variable (Framed Field Experiment)

Table A.6 shows results from GLM regressions on the remaining time. Column (1) shows a positive and statistically significant effect of the bonus treatment on remaining times. The coefficient and its standard error remains roughly unchanged with the addition of controls and fixed effects. Only in column (3) is the coefficient just barely insignificant (p-value=0.12). Column (5) shows the regression on the non-pooled framing treatments. In this specification, only the *Gain45* coefficient is significant. However, equality of coefficients of *Gain45* and *Loss45* cannot be rejected (p-value=0.16).

Table A.6: GLM regressions (student sample)

GLM: Remaining time	(1)	(2)	(3)	(4)	(5)
Bonus45 (pooled)	0.894*	0.899*	0.845	0.984*	
	(0.533)	(0.536)	(0.541)	(0.563)	
Gain45					1.252**
					(0.598)
Loss45					0.680
					(0.600)
Constant	-3.091***	-3.276	-2.574	-19.721***	-19.949***
	(0.489)	(2.018)	(2.300)	(2.297)	(2.305)
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Observations	268	268	268	268	268

Coefficients from a generalized linear model regression with a log link of the remaining time on our treatment indicators (with *Control* being the base category). Control variables added from column (2) onward include share of males in a team, a dummy whether someone in the team has been to an escape game before and average age of the team. Staff fixed effects control for the employees of *ExitTheRoom* present onsite. Robust standard errors reported in parentheses, with $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

A.5 Hint Taking at a Specific Step in the Game

We have argued that it is unlikely that hint-taking behavior alone can explain the observed performance increase of the customer teams facing incentives. In the following, we provide some additional evidence on the relationship between hint-taking and performance in our experiment. When doing so, we have to deal with two opposing effects. First, from a theoretical perspective, worse teams are more likely to use hints (which is also reflected in the positive correlation between finishing times and number of hints taken). Second, faster teams are more likely to take hints earlier on, as the likelihood of facing a difficult quest is higher for them than for slower teams. That is, if incentives make (worse) teams faster, these teams may also mechanically take more hints and this effect also accumulates over time. In order to reduce in particular the importance of the second effect, we collected information on the time at which teams reach a specific intermediate step for a subsample of 461 out of the 487 teams and compare the number of hints taken at that specific step. We focus on the point in time at which teams entered the last room of their specific task (*Zombie Apocalypse*, *The Bomb*, *Madness*), as teams reach this step on average rather early in the game. Teams facing incentives complete this step on average after 22 minutes whereas teams in the control condition need on average 24 minutes (Mann–Whitney test, p-value$= 0.018$). Hence, teams facing the incentive condi-

tion outperform control teams also early in the game. In Table A.7 we report results from ordered probit models to study whether teams facing incentives are more likely to have taken hints before the intermediate step. All five specifications reveal that team incentives do not significantly affect the number of hints taken and also none of the marginal effects of moving from one category (e.g. from one to two hints) to another category turns out to be statistically significant.

In contrast to the customer teams, we have shown that student teams (confronted with the task by us) took on average more hints when facing incentives. Repeating the analysis on reaching an the intermediate step above for the student sample shows that students facing incentives reached the intermediate step significantly earlier (they entered the last room on average after 31 minutes in *Control* and after 27 minutes when facing incentives, Mann–Whitney test, p-value$= 0.004$) but also took significantly more hints before reaching this step (see Table A.8).

Table A.7: Ordered Probit regressions: Number of hints taken when entering last room (Field Experiment)

	Ordered Probit				
	(1)	(2)	(3)	(4)	(5)
Bonus45 (pooled)	-0.018	0.012	0.113	0.050	0.134
	(0.102)	(0.105)	(0.126)	(0.185)	(0.210)
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Room Fixed Effects	No	No	No	No	Yes
Observations	461	461	461	461	461

Coefficients from an ordered Probit model. Dependent Variable: Number of hints taken at the intermediate step of entering the last room. Control variables, staff and week fixed effects as in Table 2. Robust standard errors reported in parentheses, and * $= p < 0.10$, ** $= p < 0.05$ and *** $= p < 0.01$.

Table A.8: Ordered Probit regressions: Number of hints taken when entering last room (Framed Field Experiment)

	Ordered Probit				
	(1)	(2)	(3)	(4)	(5)
Bonus45 (pooled)	0.244*	0.235*	0.285**	0.306**	0.361**
	(0.137)	(0.138)	(0.138)	(0.142)	(0.150)
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Room Fixed Effects	No	No	No	No	Yes
Observations	267	267	267	267	267

Coefficients from an ordered Probit model. Dependent Variable: Number of hints taken at the intermediate step of entering the last room. Control variables, staff and week fixed effects as in Table 7. Robust standard errors reported in parentheses, and $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

A.6 Room Fixed Effects for Natural and Framed Field Experiment

Table A.9: Probit and GLM regressions including room fixed effects

	Probit (ME): Solved in less than 45 minutes			
	Field experiment (1)-(2)		Framed Field Experiment (3)-(4)	
	Probit (ME)	GLM	Probit (ME)	GLM
	(1)	(2)	(3)	(4)
Bonus45 (pooled)	0.150***	0.266	.0763**	0.979*
	(0.056)	(0.173)	(0.038)	(0.572)
Constant		3.706***		-18.489***
		(0.511)		(1.950)
Fraction of control teams solving the task in less than 45 min	0.10		0.045	
Control Variables	Yes	Yes	Yes	Yes
Staff Fixed Effects	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes
Room Fixed Effects	Yes	Yes	Yes	Yes
Observations	487	487	268	268

The table shows average marginal effects from Probit regressions of whether a team solved the task within 45 minutes (1) and (3) and coefficients of GLM regressions on the remaining time (2) and (4) for the customer and the student sample. The specifications are as in Table 2 (1), A.2 (1), 7 (3), and A.6 (4), but include in addition Room Fixed Effects. Robust standard errors reported in parentheses, and $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.

A.7 Ordered Probit Regressions for Natural and Framed Field Experiment: Hint taking

Table A.10: Number of hints requested

	Ordered Probit: Number of hints requested							
	Field experiment (1)-(4)				Framed Field Experiment (5)-(8)			
	within 60 minutes		within 45 minutes		within 60 minutes		within 45 minutes	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bonus45 (pooled)	0.116	0.086	0.341***	0.190	0.401***	0.395***	0.878***	0.933***
	(0.094)	(0.177)	(0.094)	(0.185)	(0.140)	(0.145)	(0.142)	(0.150)
Control Variables	No	Yes	No	Yes	No	Yes	No	Yes
Staff Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Week Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Observations	487	487	487	487	268	268	268	268

Coefficients from OLS regressions of the number of hints requested within 60 minutes or 45 minutes regressed on our treatment indicator *Bonus45*. Controls and fixed effects identical to previous tables. Robust standard errors reported in parentheses, and $^{*} = p < 0.10$, $^{**} = p < 0.05$ and $^{***} = p < 0.01$.