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

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

February 14, 2022

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 3,000 participants, we document a positive effect of bonus incentives on the probability of completion of such a task. Bonus incentives seem to 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, Simon Jäger, Rajshri Jayaraman, Michael Kosfeld, Botond Kőszegi, Andreas Leibbrandt, Stephen Leider, Steven Levitt, Rocco Macchiavello, Stephan Meier, Takeshi Murooka, Susanne Neckermann, Michael Raith, Dirk Sliwka, Christian Traxler, Bertil Tungodden, Timo Vogelsang, Roberto Weber, as well as seminar participants at Augsburg, Barcelona, Bonn, Budapest, Columbia, Heidelberg, Johns Hopkins, Karlsruhe, Lausanne, Munich, NBER OrgEc Meeting, Regensburg, Tilburg, 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) & CEPR & CESifo, Geschwister-Scholl-Platz 1, D-80539 Munich.

‡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. Examples include teams of IT professionals, specialist doctors, as well as management consultants. These teams often face a series of novel and complex problems and need to gather, evaluate, and recombine information in order to succeed; frequently in a limited amount of time. Autor et al. (2003) analyze task input in the U.S. 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 their importance for later years. Given their pervasiveness 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 monetary incentives affect team production in these cognitively demanding, interactive and diverse tasks. 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., Bandiera et al., 2005, 2013; Delfgaauw et al., 2015; Englmaier et al., 2017; Erev et al., 1993; Friebel et al., 2017; Hossain and List, 2012; Jayaraman et al., 2016; Lazear, 2000; Shearer, 2004), evidence on the effects of bonus incentives is scarce for non-routine analytical tasks in which teams jointly solve a complex problem and are likely to be also intrinsically motivated.

For many modern tasks, contracts make use of performance-related bonus payments as an important part of compensation (Lazear and Oyer, 2013). Although bonus incentives appear simpler than optimal incentives prescribed by standard theories, firms frequently use bonuses instead of fully state-contingent schedules (Moynahan, 1980; Churchill et al., 1993). Furthermore, behavioral aspects such as fairness concerns (Fehr et al., 2007), overconfidence (Larkin and Leider, 2012), or loss aversion (Herweg et al.,

2010) and features germane to non-routine analytical tasks (e.g. informational asymmetries) may render binary payment schemes attractive for principals (Ulbricht, 2016). Hence, bonus incentives appear not only to be an incentive scheme of high practical relevance but also as particularly important to consider in the context of non-routine and complex tasks.

This study exploits a unique field setting to measure the effects of bonus incentives 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 quests in order to succeed (usually by escaping a room within a given time limit using a key or a numeric code). This task provides an excellent environment to study our research question, as it encompasses several elements that are prevalent in many other non-routine analytical and interactive team tasks: teams face a series of complex and novel problems, need to collect and recombine information, and have to solve analytical and cognitively demanding quests that require thinking outside the box. The task is also interactive, as 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 700 teams (with more than 3,000 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 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; Azoulay et al., 2011; Ederer and Manso, 2013; McCullers, 1978; McGraw, 1978).[1] Further, non-routine analytical tasks are more likely to be performed by people who enjoy these kinds of activities and are hence intrinsically motivated (see, e.g., Autor and Handel, 2013; Delfgaauw and Dur, 2010; Friebel and Giannetti, 2009). Extrinsic incentives could negatively affect team performance by crowding

[1]Further, incentive effects may interact with whether the task is perceived as interesting (Takahashi et al., 2016) .

out such intrinsic motivation (e.g. Deci et al., 1999; Eckartz et al., 2012; Gerhart and Fang, 2015; Hennessey and Amabile, 2010).

Recent evidence from related strands of the literature on incentives for idea creation (Gibbs et al., 2017) and creativity (e.g. Bradler et al., 2014; Charness and Grieco, 2019; Gibbs et al., 2017; Laske and Schroeder, 2016; Ramm et al., 2013), 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 joint 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 finds no effects of tournament incentives on performance in pairs but it is unclear whether this null effect is robust, as the authors achieve rather low statistical power. Another strand of literature that has identified positive effects of individual and team incentives in tasks that require mainly cognitive effort are studies on the effectiveness of teachers (Fryer et al., 2012; Muralidharan and Sundararaman, 2011). One may argue that teachers (at least sometimes) face novel and unknown problems and thus perform (at least to some extent) non-routine analytical tasks, but it remains unclear if and to what extent complementarities in individual teacher performance may be regarded as features of joint team production given that teachers usually teach different subjects and not simultaneously.

Our unique field setting allows us to substantially advance the literature on incentives for jointly solved non-routine tasks. We study the causal effect of incentives on team performance as well as on teams' willingness to explore original solutions in such a task. The setting allows us to study teams from two very distinct samples in the spirit of List (2003, 2004a,b, 2006), which also differ in their motivation. First, we conducted a series of field experiments with regular teams (customers of our cooperation partner) who were unaware of taking part in an experiment.[3] These teams self-selected into the

[2]Bradler et al. (2014), Charness and Grieco (2019), 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.

[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

task and were strongly intrinsically motivated to complete it. Second, we investigate whether our main treatment effects are also observed in a sample of students who were paid to perform the task as part of an economic experiment. These teams did not self-select into the task and were exogenously formed.[4] Using additional survey responses from the student participants, we provide some initial tentative insights on how incentives affect team organization. Finally, we discuss through which channels and effort dimensions incentives may operate using additional evidence from an expert survey, an additional laboratory experiment, and a complementary field experiment on the causal role of leadership in these tasks (Englmaier et al., 2021).

To identify the effect of providing incentives in our field experiments, 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, in line with the idea that in non-routine tasks there is an important random component in how effort translates into performance, bonus incentives did not 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. As behavioral aspects relating to reference-dependent preferences appear crucial in the context of bonus incentives (Herweg et al., 2010), we implemented the bonus incentive framed either as a gain or a loss. We find that framing plays a minor role for performance. In a similar spirit, 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 perfor-

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.

mance, whereas paying a bonus for completing the task in the regular pre-specified time of 60 minutes tended towards a positive effect. Hence, a salient reference point alone is not enough to increase performance.

Our field experiments were primarily designed to identify the causal effect of bonus incentives on team performance and to decompose the effects of bonus incentives into their respective components. However, our field data allows us also to shed some light on moderators of the incentive effect. First, we find that the bonus treatment is similarly effective for different gender and age compositions of teams (if at all, the bonus appears to be slightly more effective for teams with experienced members). Thus, it appears that targeting bonuses to certain types of teams seems less important in our task. Second, our results indicate that incentives affect team organization. Answers to our ex-post survey of the student sample suggest that incentives create demand for leadership and thus affects team performance through changes in team organization. Finally, our findings highlight that introducing incentives does not lead to a strong reduction in a team's willingness to explore original solutions for the customer teams, who self-selected into the task. However, such discouragement is apparent among student teams, which were exogenously assigned to the task and performed it as part of a study they were paid for.

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 thinking outside the box. This idea has recently been strongly promoted in the public, for instance by the best selling author Daniel Pink, in his famous TED talk with more than 19 million views and his popular book *Drive* (Pink, 2009, 2011). Our results alleviate these concerns in the context of teams jointly working on a rich and diverse non-routine analytical task. We provide novel and robust evidence that bonus incentives are a viable instrument to increase performance in such tasks. Following the approach in Maniadis et al. (2014), we show that the observed effect size in our well powered field experiment should entice readers to update their belief about a positive influence of incentives strongly upwards, even when holding rather pessimistic priors. This is also true in an additional sample of less motivated teams where we replicate the positive performance effect of incentives. Furthermore, we show that it was more likely the reward component of the bonus, and not the 45 minutes reference point which improved teams' outcomes. The latter findings

also complement recent research on non-monetary means of increasing performance (for a review of this literature see Levitt and Neckermann, 2014), in particular research referring to workers' awareness of relative performance (see e.g., Blanes i Vidal and Nossol, 2011; Azmat and Iriberri, 2010; Barankay, 2010, 2012). Finally, we add novel and interesting insights to the discussion of whether incentives discourage the exploration of new approaches. The answer to this question appears to hinge crucially on the characteristics of the underlying sample. We observe such discouragement only among the student sample, in which, presumably, less intrinsically motivated teams worked 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.

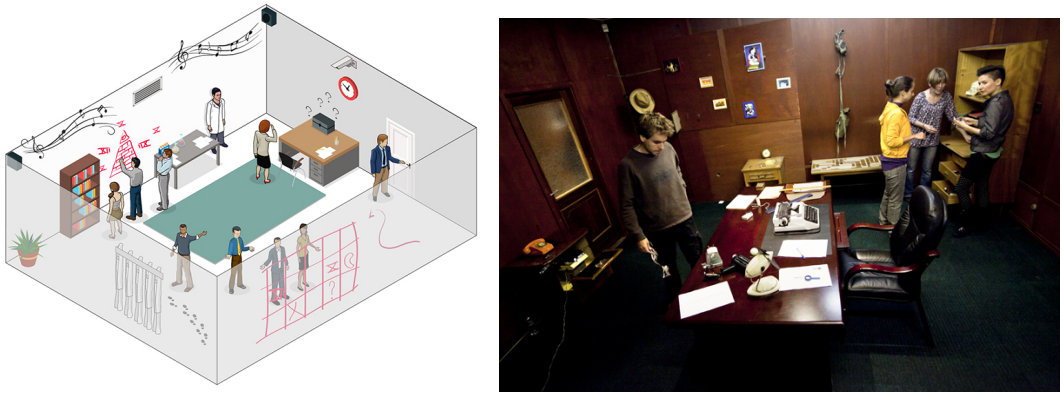
The rest of this paper is organized as follows: Section 2 presents the field setting and the experimental design. Section 3 provides the results from both field experiments. We discuss potential mechanisms in Section 4, and provide a more general discussion in Section 5. Section 6 concludes.

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 have to solve, in a real setting, a series of quests 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 complete the task they are facing within a given time limit. If a team manages to complete the task before the allotted

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



Notes: 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

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 complete the task. Typically, not all items will contain helpful information, and part of the task is determining which items 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

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

quest is to 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 Erat and Gneezy, 2016; Kachelmaier et al., 2008).

We conducted our experiments at the facilities of ETR 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 task will be declared as completed 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 incomplete and the activity ends. If a team gets stuck, they can request hints via radio from the staff at ETR. As they can only ask for a total of up to five hints, a team needs to state explicitly that they want to receive a hint. The hints never contain the direct solution to a quest, but only provide vague clues regarding the next required step.

ETR provides a rich setting with many aspects of modern non-routine analytical team tasks. First, finding clues and information very much matches the research activity that is often necessary before collaborative team work begins. Second, combining the discovered information is not trivial, and requires ability for complex 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 complete the task, the subjects must effectively cooperate as a team. As in other non-routine team tasks, team members are supposed to provide additional angles to the problem at hand, and substantial synergy effects of different approaches to problem solving will enable a team to complete the task more quickly. Fourth, 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 the activity). We interpret the fact that many teams opt to write

[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. 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. We refrain from presenting the regression specifications with room fixed effects in the main text but provide these specifications in the Appendix. Adding room fixed effects does not change our results (see Table A.1).

their names and finishing times on the walls of the entrance area of ETR as evidence for a strong motivation to finish quickly. Especially if teams are driven by the challenge of solving puzzles and take enjoyment from progressing in the task, succeeding as fast as possible is clearly desirable. Most importantly and objectively, teams never know how many intermediate quests are left to complete the task in its entirety. Hence, if a team wants to complete the task, the team has a strong incentive to succeed quickly. Finally, the team task is both difficult and non-routine in nature. This is corroborated by the fact that a substantial fraction of teams fail to finish in 60 minutes (33 percent of customer teams and 52 percent of student teams) without incentives, and a substantial fraction of teams with experienced team members (28 percent in the field experiment and 50 percent in the framed field experiment) fail to do so.[8]

The properties of these tasks are defining features of a broad class of modern jobs. Deming and Kahn (2018) find that many modern jobs require both, cognitive skills such as problem solving, research, analytical and critical thinking, as well as social skills such as communication, teamwork, and collaboration. Further, employers routinely list teamwork, collaboration, and communication skills as among the most valuable, yet hard to find qualities of workers (Deming, 2017; Casner-Lotto and Barrington, 2006; Jerald, 2009). Akin to the skills required in our escape game, employers who were asked which attributes they seek on a candidate's resume in the National Association of Colleges and Employers Survey (NACE, 2015) rank leadership skills, ability to work in a team, problem solving skills, strong work ethic and analytical and quantitative skills among the top 6.

While these features therefore 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.

[8] In the field experiment, 48 percent of customer teams have at least one experienced team member. Among the student sample, 36 percent do so. With incentives, still more than 15 percent of experienced teams fail to finish the task in 60 minutes in the field experiment and about 40 percent in the framed field experiment.

Of course, there are limitations to our setting that need to be discussed as well. First, ETR customers choose to perform the task for enjoyment and are willing to incur some costs in order to do so. This suggests that they are likely to receive some utility from performing the task (e.g. they are driven by the challenge of solving puzzles and tackling different angles of the complex task), which may not hold more generally for the choice of an occupation. However, many employees working on non-routine analytical team tasks (e.g. teams of IT specialists or specialist doctors) have also self-selected into their occupation and incurred substantial costs (e.g. in terms of education) to be able to perform interesting non-routine tasks in their job.[9] Further, we wanted to be able to study the effects of bonus incentives in a sample that provides a quite extreme benchmark in terms of intrinsic motivation, and complement results from this specific sample with results from student teams which were exogenously formed and paid by us to perform the task as part of an economic experiment. As we observe very similar effects of incentives on teams' finishing times across both samples, it seems that this particular feature (i.e. interest in performing the task) is not crucial to the effectiveness of our bonus treatment.

Second, non-routine analytical team tasks are diverse in nature. While finding and (re-)combining the discovered information, and effectively cooperating as a team are certainly distinguishing features, our task may not feature all aspects of modern teamwork. For instance, intrinsic motivation to perform these tasks (for example in business or academia) may not solely stem from making progress in and eventually completing them, but also from salient greater goals that team success can achieve. As the escape game does not feature these, it is worthwhile to discuss its implications for external validity in more detail. One could argue that a lack of such goals reduces external validity, as the effectiveness of incentives may hinge on workers' motivation. As we do find that incentives increase performance, both for people who value performing the task (customer sample) and people being hired to complete the task (student sample), it is unlikely that a lack of intrinsic motivation (due to a lack greater goals) affects our main findings. There-

[9] An intrinsic desire for being able to perform non-routine analytical jobs has been long recognized and leveraged by recruiters. One notable example are some of Google's recruiting campaigns featuring signs, placed at Harvard Square and across the Silicon Valley. These signs did not reveal to be associated with Google, but instead challenged passers-by to solve a complicated math problem. The correct answer led to a website that posed yet another puzzle. Eventually, the determined problem-solver arrived at an official Google recruiting website that asked them to submit their resume. See `https://www.npr.org/templates/story/story.php?storyId=3916173&t=1534099719379`.

fore we consider our results to be informative for a large number of work environments comprising these task elements.

Third, one could argue that in some environments there may exist more than one single solution to a complex problem, while in our setting there is only one. However, most complex problems of interest arguably have only a single (optimal) solution, but there exist multiple ways of arriving at that solution, both in the work place, as well as in our setting. More specifically, we think of incentives as trying to motivate the worker to produce the best possible solution in a given amount of time (by identifying the main problems to be solved and coming up with a solution). For example, consider a team of IT specialists that is confronted with a complex task in which they have to develop a platform that fulfills predefined requirements within a specific time frame. To this end, team members have to identify the main constraints and develop tailored solutions. While there may be several new platforms that the team can develop, most likely only one of them will be optimal given the demands by the employer (e.g. in terms of specifications or expected sales). Thus, even if several platforms can be developed, the employer will want to incentivize the team to find the optimal solution and not an inferior one.

Fourth, the fact that our subjects work in very close proximity to their team members may alleviate potential free rider concerns common to regular office settings. In the absence of free riding, we would thus estimate inflated incentive effects. However, as the task requires mainly cognitive effort, observability of co-workers' effort provision is limited in our setting. Furthermore, if the utility from completing the task quickly without contributing was lower than in a comparable work setting, we should observe differences in performance effects among highly intrinsically motivated (customer sample) and less highly intrinsically motivated teams (student sample). However, incentives increase performance in both samples similarly.

Finally, we like to note that while our task lasts much longer than usual tasks in laboratory experiments, incentives in work environments are frequently designed to stimulate effort over even longer periods, such as weeks, months or years. We deem the question of how to optimally design incentives over such time spans as very important, but clearly, our experiment was not designed to investigate long-run effects of bonus incentives. Instead we study the general effectiveness of bonus incentives in non-routine analytical team tasks in the light of widespread claims of "if-then rewards" being ineffec-

tive in such modern tasks (Pink, 2009, 2011) and provide robust evidence that incentives do improve team performance using an objective measure.

2.2 Experimental Treatments and Measures of Performance

We conducted the field experiment with 3308 customers (722 teams) 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 team bonus if they managed to complete 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 complete 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 complete the task in less than 45 minutes.[10]

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 teams), we explicitly mentioned the 45 minutes as a salient reference point before the team started working on the task. However, we 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 teams) and *Loss60* (46 teams), we provided a monetary bonus but did not provide the reference point of 45 minutes: Teams received the bonus if they completed the task within 60 minutes.

[10]The bonus amounted, on average, to approximately €10 per team member. Teams in the field experiment 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 for the team to earn 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. In order to preserve the natural field experiment, we always made sure that the experimenters blended in with the ETR staff.

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[11], team language (German or English), experience with escape games, and whether the customers came as a private group or were part of a company team building event[12]. Our primary outcome variable is team performance, which we measure by i) whether or not teams completed 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 a benefit from providing incentives in a loss frame compared to a gain 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.[13]

Further, we replicated our main treatments (*Gain45*, *Loss45* and *Control*) in a framed field experiment at ETR in which we randomly allocated student participants from the subject pool of the social sciences laboratory at the University of Munich (MELESSA) into teams (804 participants in 268 teams). The additional sample allows us to study whether bonuses affect team performance in similar ways when we form teams exogenously and pay them to perform the task as part of an economic experiment. Further, it enables us to collect additional data on task perception and team organization.

[11] Again, note that to preserve the natural field experiment, we did not interfere with the standard 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. As requested by the IRB, teams with minors were not included in the study.

[12] ETR staff regularly ask teams whether they have ever participated in an escape game and whether the nature of the group is private or a team building event irrespective of our experiment.

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

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.[14] 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*. We randomized on a daily level to avoid treatment spillovers between different teams on-site (as participants from one 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.[15]

Upon arrival, ETR 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 working on 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

[14] ETR 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.

[15] 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.

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.[16]

Upon arrival at the laboratory, the participants were informed about their upcoming participation in an escape game. They 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 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 essay.[17] 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 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 we made sure that student teams in different sessions on a given day did not encounter 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,

[16] Section A.3 in the Appendix provides a translation of the text of the invitation.

[17] 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.

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.[18]

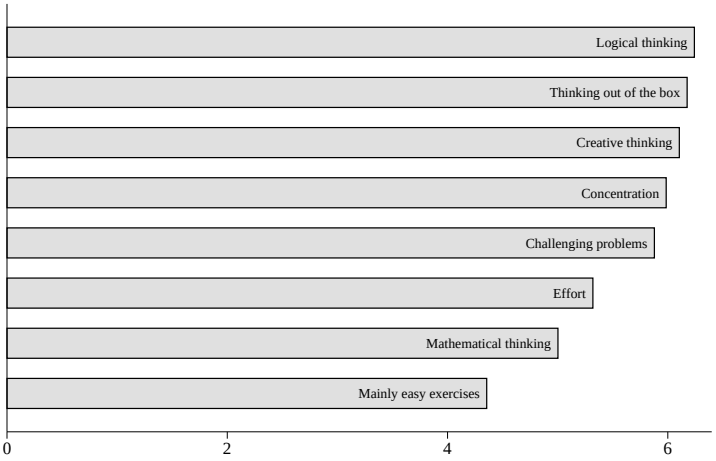
3 Results

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 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 then explore 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. 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 encompass many features of modern non-routine analytical tasks as teams face novel and challenging problems that require cognitive effort, analytical thinking and thinking outside the box rather than easy repetitive chores. In order to not interfere with the standard procedures at ETR, 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).

[18] 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 earning 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.



Notes: 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

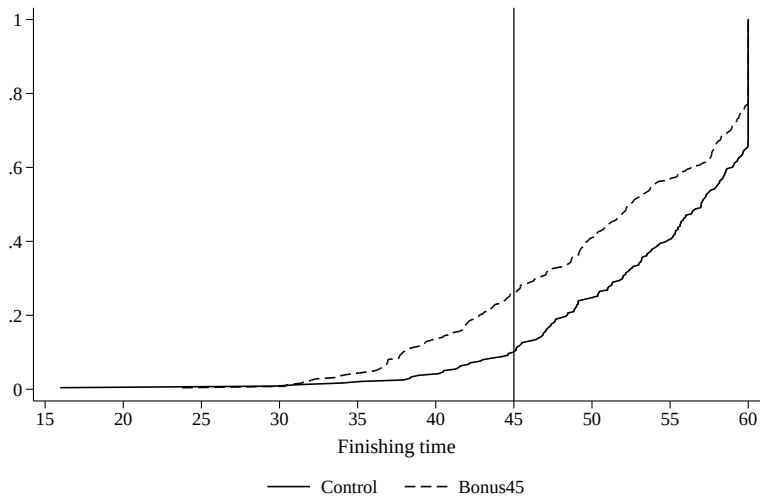
Figure 2 shows the mean answers of our participants. Participants strongly agreed that the task involves logical thinking, thinking outside the box, and creative thinking, in particular as compared to mathematical thinking and easy exercises (signed-rank tests reject that the ratings have the same underlying 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* $p = 0.02$).

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 English-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), also when adjusting for multiple hypothesis testing according to List et al. (2019). 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).

Table 1: Sample size and characteristics

	Control (n=238)	*Bonus45 (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}***

Notes: 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), and *** p<0.01, ** p<0.05, * p<0.1. Adjusting for multiple hypotheses testing according to the procedures in List et al. (2019) does not affect the significance of p-values indicated above.



Notes: 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

3.2 Bonus Incentives and Team Performance

We now turn to our primary research question, whether providing bonus incentives improves performance. As mentioned earlier, our objective outcome measure of performance is whether teams manage to complete the task within 45 minutes and more generally how much time teams need. 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. In line with the idea that non-routine tasks are characterized by a stochastic process which translates effort into performance, we observe differences over a large part of the support of the distribution rather than merely around the 45 minutes threshold. In *Control*, only 10 percent of the teams manage to finish 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 completion also differs significantly between *Bonus45* and *Control* (p-value < 0.01, Mann–Whitney test). In *Bonus45*, 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). Adjusting p-values for multiple hypotheses testing as suggested in List et al. (2019) yields similar results (see Section A.4.1 in the Appendix for more details).

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 completing the task within 45 minutes. We cluster standard errors at the day level (at which we varied the treatment) throughout.

Column (1) includes only a dummy variable for the bonus treatments *Bonus45*. Bonus incentives are estimated to increase the probability of completing the task in less than 45 minutes by 16.5 percentage points. In Column (2), we add observable team characteristics. Here, and in the following analysis, group size, experience with escape games and the share of males in a team have a positive effect on performance whereas English-

Table 2: Probit regressions: Completed in less than 45 minutes

	Probit (ME): Completed in less than 45 minutes				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.165***	0.164***	0.188***	0.151***	
	(0.024)	(0.022)	(0.025)	(0.041)	
Gain45					0.125***
					(0.037)
Loss45					0.174***
					(0.046)
Fraction of control teams completing 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

Notes: The table displays average marginal effects from probit regressions of whether a team completed the task within 45 minutes on our treatment indicators (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 the group speaks German and a dummy for private teams (opposed to company team building events). Staff fixed effects control for the employees of ETR present on-site 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.6 in Section A.5 of the Appendix reports regressions from a sample excluding weeks without variation in the outcome variable). Robust standard errors clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

speaking groups perform slightly worse.[19] 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 completing a non-routine task strongly enhances their performance. We also estimate the effects of bonuses on the time remaining upon completion of the task, which confirm both the results from the non-parametric tests on the remaining time as well as the results from the probit models in Table 2 (see Table A.7 in Appendix A.5.3).

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 lose its effect in the last 15 minutes. In addition, if incentives crowd out intrinsic motivation to provide effort, 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

[19]See also Table A.5 in the Appendix. Table A.5 further shows that the treatment effect does not strongly interact with the observable team characteristics. Only the interaction of incentives and experience (model (4) in Table A.5) turns out to be significantly positive at the five percent level.

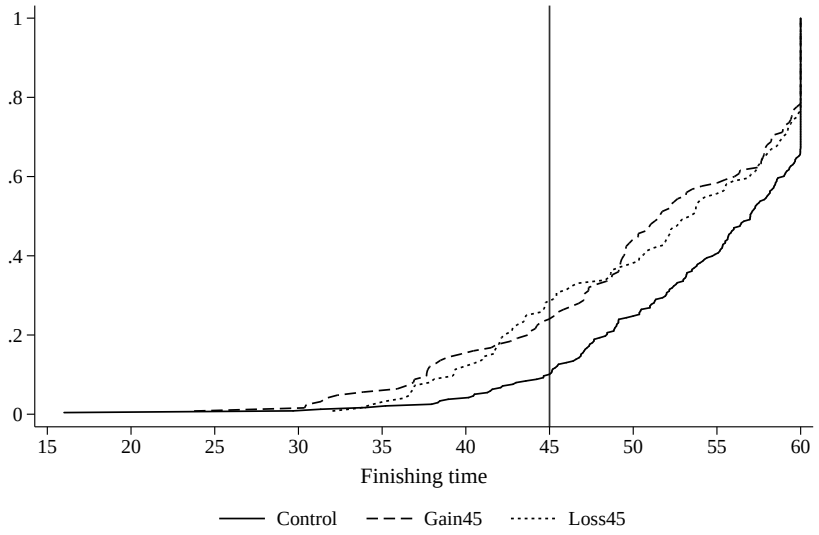
Table 3: Influence of main bonus treatment on hazard rates

	Cox Proportional Hazard Model: Finishing the Task					
	First 45 min (1)-(3)			Last 15 min (4)-(6)		
	(1)	(2)	(3)	(4)	(5)	(6)
Bonus45	2.853***	2.947***	2.914***	1.178	1.251	0.841
	(0.446)	(0.474)	(0.844)	(0.189)	(0.248)	(0.180)
p-value for prop. haz. assumption	0.830	0.748	1.000	0.800	0.686	0.995
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

Notes: 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 clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1. 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.

prior was true, we should observe the treatment to have a strong effect on the hazard in the first 45 minutes, and no or even a negative effect in the last 15 minutes, conditional on covariates.

Table 3 shows the hazard ratios using our usual set of controls and employing cluster-robust standard errors. Columns (1) through (3) estimate the effect on the hazard rate for the first 45 minutes and columns (4) through (6) for 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 ranges closely around one and is not significantly different from one in any specification. 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



Notes: 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 *Gain45*, *Loss45*, and *Control* in the field experiment

out intrinsic motivation in our setting (after teams fail to achieve the bonus payment). We conclude:

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

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. Participants arrived at the facility not expecting any payment at all, therefore both frames have the same absolute distance from a reference point of zero.[20] Figure 4 shows the cumulative distributions of finishing times for both frames separately.

We find that the framing of the bonus appears to be of minor importance for team performance. A Mann–Whitney test fails to reject the null hypothesis that the finishing

[20]It seems unlikely that participants were forming any other reference point than zero. Payment for the activity was usually done weeks in advance through the company's website and should therefore not affect reference points when entering the facility at a much later date.

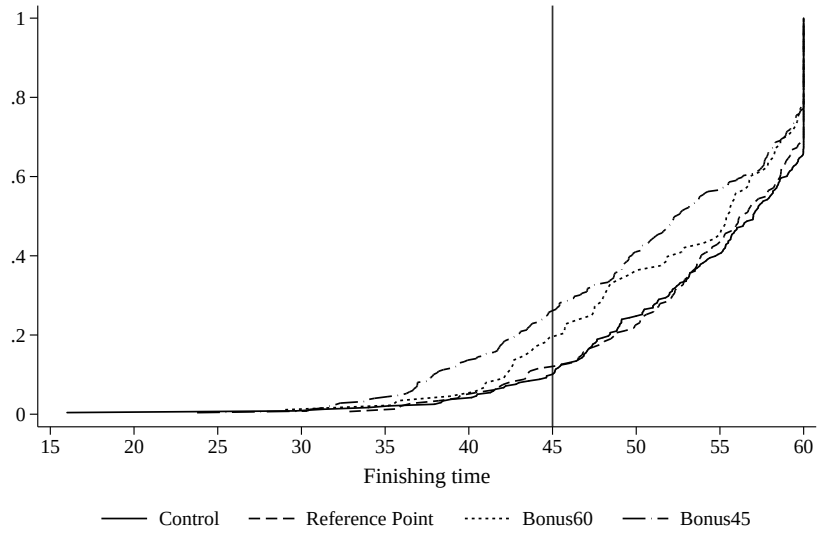
Table 4: Task performance for main treatments

	Control	*Bonus45*	*Gain45*	*Loss45*
Fraction of teams completing task in 45 min	0.10	0.26***	0.24***	0.28***
Fraction of teams completing task in 60 min	0.67	0.77**	0.78**	0.77*
Mean remaining time (in sec)	345	530***	548***	512***
Mean remaining time (in sec) if completed	515	688***	707***	669***

Notes: This table summarizes key variables and their differences across our three treatments *Control*, *Gain45*, and *Loss45*, and the pooled bonus incentive treatment (*Bonus45*). Stars indicate significant differences from *Control* (using χ^2 tests for frequencies and Mann–Whitney tests for distributions), and *** p<0.01, ** p<0.05, * p<0.1. See Table A.2 in the Appendix for MHT adjusted p-values according to List et al. (2019).

times for the two framings come from the same underlying distribution (p-value $= 0.70$). Also, the fractions of teams completing 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 completing 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 and Appendix Table A.2 highlights that the observed incentive effect is robust to controlling for multiple hypotheses testing using procedures recommended in List et al. (2019).

In addition to the non-parametric analyses, we report results from a regression of the probability of completing 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 completing the task within 45 minutes under both frames (as compared to the control condition). The average marginal effect for the *Loss45* treatment is estimated to be 5 percentage points larger and a post-estimation Wald test for the equivalence of the coefficients *Gain45* and *Loss45* in Column (5) of Table 2 identifies a statistically significant difference across the two frames (Wald test, p-value < 0.05). However, the same test fails to achieve significance at the ten percent level in alternative specifications that either exclude staff and week fixed effects (Wald test, p-value $= 0.26$) or use Huber-White standard errors instead of clustering standard errors at the day level (Wald test, p-value $= 0.38$). Furthermore, the results in Table A.7 show



Notes: The figure shows the cumulative distribution of finishing times of *Bonus45* (pooled), *Bonus60* (pooled), *Reference Point* and *Control*. The vertical line marks the time limit for the bonus in the *Bonus45* condition.

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

that framing bonuses as losses does not seem to have any additional effect on the time remaining (Wald test, p-value $= 0.98$). We thus summarize our findings as follows in Result 2.

Result 2 *The framing of bonuses plays a minor role.*

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 conducted 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 completing the task in 60 minutes.[21] 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*,

[21] We do not differentiate between the gain and the loss frame of *Bonus60* in the following. As for most of the analysis with respect to *Bonus45*, no difference between the frames in *Bonus60* emerged.

Table 5: Probit regressions: Completed in less than 45 minutes (all treatments)

	Probit (ME): Completed in less than 45 minutes			
	(1)	(2)	(3)	(4)
Bonus45	0.160***	0.157***	0.164***	0.108***
	(0.023)	(0.022)	(0.026)	(0.035)
Bonus60	0.105**	0.102***	0.105***	0.127**
	(0.041)	(0.038)	(0.039)	(0.051)
Reference Point	0.025	0.023	0.011	0.020
	(0.032)	(0.035)	(0.039)	(0.039)
Fraction of control teams completing 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

Notes: The table shows average marginal effects from probit regressions of whether a team completed the task within 45 minutes on our treatment indicators *Bonus45* (pooled), *Bonus60* (pooled), and *Reference Point* with *Control* being the base category. Control variables, staff and week fixed effects as in Table 2. Robust standard errors clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

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 $= 0.78$). The results point in a similar direction when adjusting for multiple hypothesis testing following the approach suggested in List et al. (2019); details are presented in Appendix A.4.1.

Lastly, we provide a regression analysis for the full sample of ETR customer teams 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.[22] 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 (models (1) to (4), p-values < 0.1). Similarly, the coefficient of *Bonus45* is significantly

[22]Table A.8 in Appendix A.5 confirms these findings for remaining time as dependent variable.

larger than the coefficient of *Reference Point* in all specifications (p-value $= 0.07$ in model (4) and p-value < 0.01 in all other specifications). Equality of coefficients of *Bonus60* and *Bonus45* can only be rejected for one of the specifications (model (2), p-value $= 0.095$). We summarize this finding in Result 3:

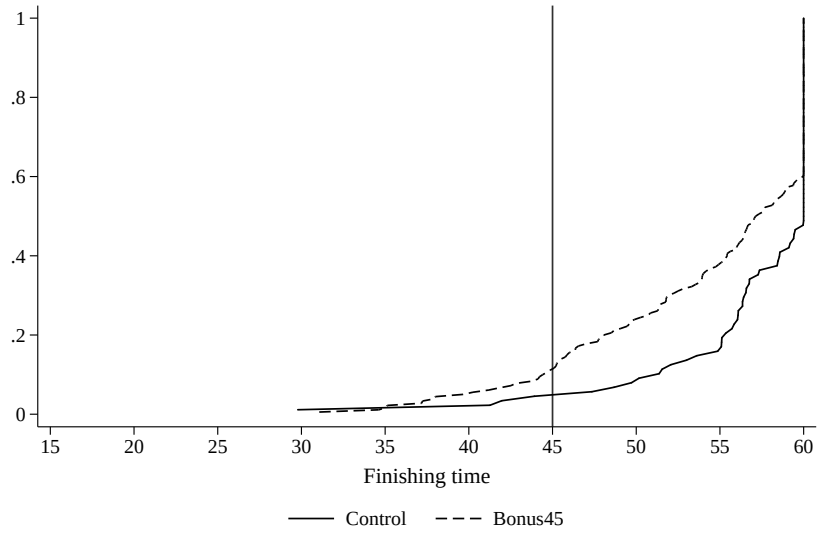
Result 3 *Bonuses increase performance due to the reward they provide. Introducing a salient reference performance (indicating extraordinary performance) is not sufficient to induce a performance shift.*

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

We have shown that bonus incentives increase performance in a sample of self-selected and motivated teams of ETR customers. To test whether the performance-enhancing effect of bonus incentives is also present in teams 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, student teams from the laboratory subject pool are confronted by us with the task, do not pay for it (but instead are paid to perform it as part of an economic experiment), and hence are less likely to be intrinsically motivated to complete the task. Teams in the field experiment are also formed endogenously and vary in size, whereas we randomly assign students to teams of three participants. Finally, our student participants differ along several observable dimensions, such as age, gender, and experience with the task.[23]

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, the average share of males in *Gain45* (0.39) tends to be lower as compared to *Control* (0.46) (Mann–Whitney test, *Gain45* vs. *Control*, p-value $= 0.08$) and *Loss45* (0.47) (Mann–Whitney test, *Loss45* vs. *Gain45* p-value $= 0.10$, *Loss45* vs. *Control*,

[23]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).



Notes: 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 in *Bonus45* and *Control* in the framed field experiment (student sample)

p-value $= 0.97$), and the share of teams with at least one team member with experience in escape games appears to be 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. *Gain45*, p-value $= 0.38$). These differences are not very pronounced and adjusted p-values according to List et al. (2019) are all larger than 0.10. Nevertheless, we control for team characteristics 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 completing the task in 45 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 on average substantially worse 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 complete 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

Table 6: Task performance for main treatments (student sample)

	Control	*Bonus45*	*Gain45*	*Loss45*
Fraction of teams completing task in 45 min	0.05	0.11*	0.13**	0.09
Fraction of teams completing task in 60 min	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 completed	355.98	546.62***	590.10**	510.50***

Notes: This table summarizes key variables and their differences across our three treatments *Control*, *Gain45* and *Loss45*, as well as the combined *Bonus45* (pooled) for the student sample. Stars indicate significant differences from *Control* (using χ^2 test for frequencies and Mann–Whitney tests for distributions), and *** p<0.01, ** p<0.05, * p<0.1. P-values of non-parametric comparisons between *Gain45* and *Loss45* exceed 0.10 for all four performance measures.

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. These results are also robust to adjusting p-values for multiple hypotheses testing as suggested in List et al. (2019) (see Section A.4.2 in the Appendix for more details).

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.[24] Table 7 reports the results from probit regressions on the probability of completing the task within 45 minutes. Column (1) only uses the treatment dummy and shows that bonus incentives significantly increase the probability of completing the task in 45 minutes. The positive effect of the bonus incentives is 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. However, a post-estimation Wald test cannot reject equivalence for

[24] 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.

Table 7: Probit regressions: Completed in less than 45 minutes (student sample)

	Probit (ME): Completed in less than 45 minutes				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.075*	0.073*	0.075*	0.079**	
	(0.042)	(0.041)	(0.039)	(0.036)	
Gain45					0.101**
					(0.039)
Loss45					0.051
					(0.041)
Fraction of control teams completing the task in less than 45 min	0.05	0.05	0.05	0.05	0.05
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

Notes: The table shows average marginal effects from probit regressions of whether a team completed the task within 45 minutes on our treatment indicators (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 ETR present on-site 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.10 in Section A.6 of the Appendix reports regressions from a sample excluding weeks without variation in the outcome variable). Robust standard errors clustered at the session level reported in parentheses, and *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

the coefficients of *Gain45* and *Loss45* at the ten percent level. Also for the student sample, the positive effect of bonus incentives is reflected qualitatively in the analyses of the time remaining (see Table A.9 in Appendix A.6).

3.5 Bonus Incentives and the Willingness to Explore

The effectiveness of bonus incentives may depend on whether monetary incentives crowd out intrinsic motivation to explore original solutions, 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

Table 8: Hints requested in the field experiment and the framed field experiment

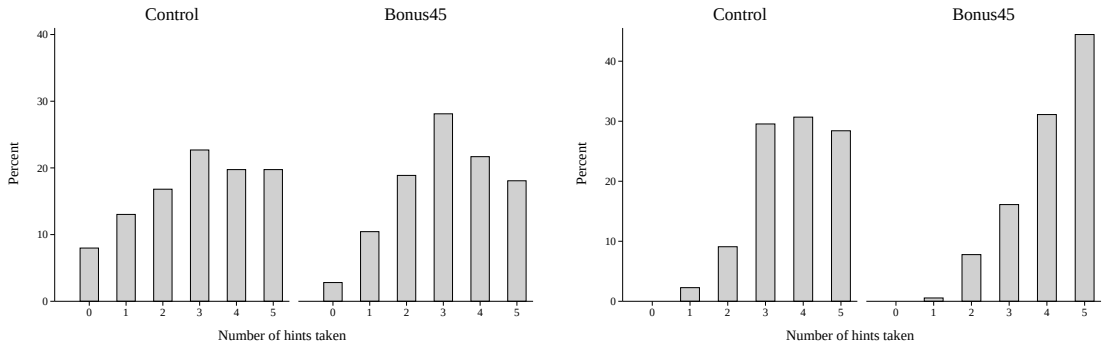
	Control	*Bonus45*	*Gain45*	*Loss45*
within 60 minutes				
Field Experiment (487 groups)	2.92(1.55)	3.10(1.34)	3.05(1.40)	3.15(1.29)
Framed Field Experiment (268 groups)	3.74(1.04)	4.11(0.98)***	4.10(0.98)**	4.12(0.98)**
within 45 minutes				
Field Experiment (487 groups)	1.97(1.22)	2.36(1.15)***	2.30(1.19)**	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)***

Notes: 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), and *** p<0.01, ** p<0.05, * p<0.1. *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.

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.[25]

Table 8 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

[25]In section A.9, we provide additional analyses showing that the increase in hint taking in the framed field experiment is unlikely due to increase importance of risk aversion when incentives are in place.



(a) Customer Sample (487 groups)

(b) Student Sample (268 groups)

Notes: 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

Control. Regression analyses for hint taking including additional controls (see Table 9, models (1), (2), (5), and (6)) confirm these results.[26]

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 9). When we add additional controls and fixed effects (columns (4) and (8) of Table 9), the results for the student sample remain largely unchanged, whereas the positive coefficient of the incentive condition becomes smaller and statistically insignificant in the customer sample.

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 the task are presumably more intrinsically motivated to work on the task and may derive utility from making progress on their own. Thus, they are 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-

[26] An ordered probit regression yields qualitatively similar results, see Table A.11 in the Appendix.

Table 9: OLS regressions: Number of hints requested

	OLS: Number of hints requested							
	Field experiment (1)-(4)				Framed Field Experiment (5)-(8)			
	within 60 min		within 45 min		within 60 min		within 45 min	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bonus45	0.172	0.098	0.387**	0.186	0.372**	0.343**	0.843***	0.808***
	(0.167)	(0.183)	(0.152)	(0.134)	(0.145)	(0.136)	(0.128)	(0.122)
Constant	2.924***	4.037***	1.971***	1.770***	3.739***	5.449***	2.330***	4.236***
	(0.130)	(0.442)	(0.109)	(0.469)	(0.126)	(0.650)	(0.102)	(0.698)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Staff FE	No	Yes	No	Yes	No	Yes	No	Yes
Week FE	No	Yes	No	Yes	No	Yes	No	Yes
Observations	487	487	487	487	268	268	268	268

Notes: Coefficients from OLS regressions of the number of hints requested within 60 minutes or 45 minutes regressed on our treatment indicator *Bonus45* (pooled). Controls and fixed effects (FE) identical to previous tables. Robust standard errors clustered at the day (for the field experiment) or session (for the framed field experiment) level reported in parentheses, and *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

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 obvious 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 strongly reduce exploration behavior by teams hired to perform the task (student teams) but affect exploration behavior of teams choosing to perform the task (customer teams), if at all, to a much smaller extent.*

4 Mechanisms

The previous results have shown that incentives causally and unambiguously improve team performance. To investigate how teams' behavior changes, we employ two strategies. First, we surveyed student teams after they performed the task. The survey results highlight that incentives may alter team organization and increase the demand for leadership as well as the endogenous emergence of a team leader. We discuss this important mechanism in light of recent evidence from a field experiment that identifies the causal effect of leadership in the same setting (see Englmaier et al., 2021). Second, we

provide a broader discussion of the dimensions along which incentives may change behavior within teams. To do so, we collected expert beliefs to identify which behavioral dimensions might be most strongly affected by incentives. While it is beyond the scope of this paper to differentiate experimentally between all possible explanations, we seek to provide some guidance in terms of promising avenues for future research through an additional smaller scale laboratory experiment, experimentally testing whether (the top) three dimensions experts considered most important were substantially affected by incentives.

4.1 Performance and Team Organization

We conducted two post-experimental questionnaires in our student sample to analyze potential mechanisms through which the treatment effect could operate. In Questionnaire 1, we asked 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 student participants), we use an additional set of questions based on the concept of team work quality by Hoegl and Gemuenden (2001). Table 10 shows the results from Questionnaires 1 and 2, reporting uncorrected p-values, as well as p-values adjusted for multiple hypothesis testing with 31 outcomes following List et al. (2019).

The upper panel of Table 10 shows that incentives in general do not strongly affect agreement with the statements we provided. However, it reveals some interesting insights about the channels through which incentives might potentially operate. First, teams appear to be notably more stressed when facing incentives than teams in *Control* (Mann–Whitney test, p-value < 0.01).[27] 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 $= 0.88/0.99$), suggesting that incentives caused positive rather than negative stress among the team members. Second, participants in the incentive treatment tend to agree more with the statement that "one team member was dominant in leading the team" (Mann–Whitney test, p-value $= 0.03/0.40$), and also with the statement "I was dominant in leading the team" (Mann–Whitney test, p-

[27] We are agnostic about whether this increase in stress levels is a direct result of incentives or a byproduct of increased effort levels.

Table 10: Answers to post-experiment questionnaires

	Control	*Bonus45*	p-value / MHT adjusted
Questionnaire 1 (n=804)			
"The team was very stressed."	3.57	4.13***/†††	0.000 / 0.000
"One person was dominant in leading the team."	2.60	2.86**	0.028 / 0.396
"We wrote down all numbers we found."	5.64	5.50**	0.044 / 0.991
"I was dominant in leading the team."	2.64	2.87**	0.053 / 0.520
"We first searched for clues before combining them."	4.58	4.39	0.107 / 0.899
"We exchanged many ideas in the team."	5.87	5.74	0.119 / 0.904
"When we got stuck we let as many team members try as possible."	5.43	5.28	0.143 / 0.914
"The team was very motivated."	6.14	6.26	0.221 / 0.881
"We communicated a lot."	5.78	5.88	0.227 / 0.982
"All team members exerted effort."	6.23	6.37	0.242 / 0.850
"Our notes were helpful in finding the solution."	5.50	5.43	0.413 / 0.999
"I was able to present all my ideas to the group."	5.95	5.93	0.406 / 0.991
"We were well coordinated in the group."	5.73	5.80	0.606 / 0.997
"I was too concentrated on my own part."	2.88	2.83	0.763 / 1
"We made our decisions collectively."	5.51	5.58	0.867 / .999
"I would like to perform a similar task again."	6.30	6.28	0.876 / 0.985
"Our individual skills complemented well."	5.65	5.68	0.891 / 0.998
"The mood in our team was good."	6.30	6.36	0.929 / 0.992
"All team members contributed equally."	5.97	6.00	0.956 / 0.999
Questionnaire 2 (n=375)			
"How much did you wish somebody would take the lead?"	2.67	3.32***/†††	0.000 / 0.009
"How well led was the team?"	3.85	4.21**	0.036 / 0.400
"How much did you think about the problems?"	6.00	5.79	0.111 / 0.552
"How much did you follow ideas that were not promising?"	5.02	4.79	0.173 / 0.772
"How much team spirit evolved?"	5.54	5.80	0.168 / 0.760
"How much coordination was there of individual tasks and joint strategy?"	3.28	3.51	0.183 / 0.914
"How much exploitation was there of individual potential?"	5.14	4.94	0.217 / 0.890
"How much helping was there when somebody stuck?"	5.70	5.58	0.217 / 0.994
"How much did you search the room for solutions?"	6.31	6.22	0.515 / 0.994
"How much exertion of effort was there by all the members?"	5.98	5.96	0.600 / 0.908
"How much communication was there about procedures?"	5.30	5.35	0.883 / 1
"How much was there of accepting the help of others?"	5.80	5.85	0.892 / 1

Notes: 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, and *** p<0.01, ** p<0.05, * p<0.1. Daggers indicate significant differences when adjusting for multiple hypotheses testing (concerning 31 outcomes) according to List et al. (2019), where ††† <0.01, †† <0.05, † <0.10.

value $= 0.05/0.52$), although both of these statements miss statistical significance when adjusting for multiple hypothesis testing.

The results from Questionnaire 2 in the lower panel of Table 10 mirror the answers from Questionnaire 1. Teams facing incentives wish for more leadership (Mann–Whitney test, p-value < 0.01), while they also tend to report that teams were better led (Mann–Whitney test, p-value$= 0.04/0.40$), although the latter fails to reach conventional significance levels when adjusting for multiple hypothesis testing. Overall, both questionnaires hint at the fact that incentives may change the way how teams are organized and suggest that incentives may lead to an endogenous emergence of (a demand for) team leaders. In line with the idea that leadership may improve team performance in our context, Englmaier et al. (2021) study the value of such endogenous leadership in the same setting. They find that randomly encouraging teams to chose a leader (i.e., randomly creating demand for leadership) indeed causally improves team performance. Treated teams in their experiment are significantly more likely to complete the task (the share of teams completing the task within 60 minutes increased from 44% in *Control* to 63%) and teams in the treatment condition also completed the task considerably faster. It seems thus conceivable that incentives led to a similar change in team organization resulting in the observed performance improvements.

4.2 Incentives and Effort Dimensions

In addition to highlighting (the demand for) leadership as a central mechanism of how incentives improve team performance, we also explore which effort dimensions incentives may affect in non-routine team tasks. Based on valuable comments in seminars, workshops, conference presentations, and a brain storming among the research team, we compiled a list of ten potentially important effort dimensions (see Table 11) through which incentives may impact team performance. We then recruited experts with knowledge of behavioral and experimental economics, as well as personnel and organizational economics for participation in an online survey to consider the relative importance of incentives for each of these dimensions. [28] We contacted 104 academic economists whom we identified as working on the role of incentives in the workplace, being broadly concerned with studying the effects of (financial) incentives, or contributing to the field of

[28] The entire design, timing and intended analysis of the survey was pre-registered. For details see `https://aspredicted.org/hc8r7.pdf`.

personnel economics (if we deemed their work relevant to the present study). In January 2020, these experts received an email containing a link inviting them to fill in the survey (henceforth the expert sample). A few days later, we also sent the invitation to the discussion mailing list of the Economic Science Association (ESA-discuss) using a different link and thus generating results from a second sample consisting mostly of researchers active in behavioral and experimental economics (henceforth the ESA sample). Survey participants could rank the ten possible effort dimensions from most to least affected by incentives and add additional dimensions, if they wished so.[29] Apart from the evaluation of the relative importance of the ten different effort dimensions, the survey contained questions on participant's beliefs regarding the effectiveness of incentives (and their framing) on performance in different types of tasks, respondents' knowledge of the present paper (and some related research), as well as whether they conduct(ed) experiments on incentives themselves and their academic seniority.[30]

We received 39 responses from the expert sample and 121 from the ESA sample. In line with our pre-registration, we eliminate respondents who took less than 60 seconds, suggesting they did not take the survey carefully. We also removed those who did not rank all dimensions, leaving us with 28 and 65 responses respectively.

Table 11 shows the ten statements and their average rank of each statement across our two samples, as well as the number of wins of each statement in pairwise comparisons with the other statements. As the results show, respondents in both samples strongly agreed on the relative importance of the three statements listed at the top: *"With incentives, teams communicate more (or less)"*, *"With incentives, teams share information better (or worse) among members"*, and *"With incentives, teams select the most skilled person for a specific problem"*. In both samples, these three dimensions rank among the top 4 and win in at least 6 pairwise comparisons. For dimensions that experts rank top 4–6, there is somewhat less consensus. While both experts and ESA members expect that incentives to some extent matter for the likelihood of team members taking the initiative *"With incentives, team members are more (or less) likely to take the initiative and lead the team"* (rank 4 for experts and rank 5 for the ESA sample), experts consider incentive effects for joint problem solving (*"With incentives, team members spend more (or less) time working jointly"*, and concentration (*"With incentives, teams are more (or less) likely to give in to*

[29] None of the respondents did recommend any additional effort dimension to be considered.

[30] The survey is still accessible through `https://tilburgss.co1.qualtrics.com/jfe/form/SV_9Y3lZbt8dj9WEn3`.

Table 11: Survey results

Statement	Average Rank		# of Wins in Pairwise Comparisons	
With incentives, ...	Experts	ESA	Experts	ESA
...teams communicate more (or less)	3.54	4.52	9	7
...teams share information better (or worse) among members	4.00	4.92	8	6
...teams select the most skilled person for a specific problem	4.68	4.38	7	8
...team members are more (or less) likely to take the initiative to lead the team	4.68	5.40	6	4
...team members spend more (or less) time working jointly on a specific problem (as opposed to individually)	5.25	5.51	4	2
...teams are more (or less) likely to give in to distractions	5.50	4.54	2	8
...teams select the most confident person for a specific problem	5.57	5.57	4	3
...teams allocate more (or less) time on information search relative to problem solving	5.93	5.28	3	4
...teams allocate effort more (or less) unevenly across stages of the task	6.00	6.02	1	1
...teams think more (or less) outside the box	7.25	6.57	0	0

Notes: This table reports how our sample of experts and the sample of respondents on the ESA discuss mailing list ranks the different dimensions of team production which incentives can affect. Average rank reports the average rank assigned to a statement (from 1 to 10) across all respondents within the respective sample (i.e. the lower the average rank, the more important deem respondents this dimension). # of wins in pairwise comparisons indicates how many other statements will lose in a pairwise comparison (round-robin tournament) in the respective sample (i.e. the higher the number, the more important deem respondents this dimension).

distractions") relatively more important than ESA respondents. Vice-versa, ESA respondents consider incentive effects for concentration and for how time is spent (*"With incentives, teams allocate more (or less) time on information search relative to problem solving"*) relatively more relevant than the effects of incentives on joint problem solving. Finally, respondents in both samples consider the role of incentives relatively unimportant for effort provision across time (*"With incentives, teams allocate effort more (or less) unevenly across stages of the task"*) and do not expect that *"with incentives, teams allocate more (or less) time on information search relative to problem solving"*.

As part of the survey pre-registration, we committed to perform a small scale laboratory experiment with a non-routine team task mimicking the real-life escape room challenge. This task was tailored to test how incentives affect the three effort dimensions survey respondents ranked as most important. Following several delays due to COVID-19, we eventually implemented the laboratory experiment in Munich and Tilburg in August and September 2021 (under the locally applicable COVID-19 restrictions).[31] The non-

[31]For details on our pre-registered laboratory experiment see also `https://www.socialsciceregistry.org/trials/8073`.

routine task we used in the laboratory experiment was a modified board game version of a real-life escape challenge. We designed this non-routine task to test whether incentives causally increase i) the likelihood that teams assign the most skilled team member to a specific sub-task (*skill-to-task matching*), ii) the likelihood of team members sharing relevant information (*information sharing*), and iii) the likelihood that teams communicate more (or less, *communication*).[32] Akin to our field experiments, we implemented three treatments in a between-subjects design in the laboratory setting. In *BGControl*, teams were not provided with any monetary incentives and had 60 minutes to complete the escape board game (BG) for a flat payment of €7.50 (irrespective of success). In *BGIncentive45*, subjects could receive an additional team bonus of €30 (again framed either as a gain or loss) if they succeeded in completing the board game task in less than 45 minutes. We planned to collect data for a total of 120 teams composed of 3 individuals (40 teams in *BGControl* and 40 for each framing in *BGIncentive45*). After removing three observations from sessions in which the experimental software malfunctioned and four observations from sessions conducted by a research assistant who did not administer the treatments correctly (and about whom participants complained not to have understood the instructions), our final sample comprises of 119 independent observations.

The prevailing COVID-19 regulations affected our experiment in terms of recruitment possibilities, physical distancing, and hygiene measures. All of these may have negatively influenced finishing times and difficulty as compared to the real-life escape games in our field experiments (which were conducted before the pandemic). The fraction of teams solving the task within 60 minutes in the laboratory task amounts to only 35 percent (*BGIncentive45*: 33 percent, *BGControl*: 39 percent, χ^2 test p-value $= 0.49$), which is substantially lower than in our natural field experiment (72 percent) and our framed field experiment (56 percent). Focusing on primary outcomes that were directly or indirectly incentivized by the bonus condition (i.e., remaining times and task completion within the bonus target), we nevertheless observe a tendency that teams perform better in the bonus condition: Teams' average remaining times amount to 203 seconds in *BGIncentive45* versus 174 seconds in *BGControl* and incentives tend to also increase the fraction of teams

[32]For i), we instructed teams to select one team member for a subtask requiring logical reasoning. All team members performed a logical reasoning test with individual feedback at the beginning of the experiment. For ii), we distributed a unique envelope containing information to each individual team member, whose content could be shared at a later stage. For iii), we elicited how much teams communicated (see Appendix A.10 details on the additional laboratory experiment.

solving the task within the incentive target of 45 minutes (*BGIncentive45*: 7 percent, *BGControl*: 2 percent). Due to substantial noise in the data, which was possibly amplified by adverse effects resulting from COVID-19 measures, these tentative results fail to be statistically significant (Mann-Whitney test for remaining times: p-value $= 0.81$, χ^2 test for fraction of teams completing the task within 45 minutes: p-value $= 0.26$). However, incentives do statistically significantly improve remaining times among teams who finish the task (617 seconds remaining in *BGIncentive45* versus 444 seconds in *BGControl*, Mann-Whitney test, p-value $= 0.088$), indicating that the bonus incentive is particularly effective among teams that are more also likely to achieve the bonus target.

Focusing on how incentives affect the three effort dimensions our survey respondents considered most important, we cannot reject that teams share information similarly with and without incentives (on average 1.73 members share information in *BGIncentive45* (std. dev.: 1.47) versus 1.72 members do so in *BGControl* (std. dev.: 1.46), Mann-Whitney test, p-value $= 0.97$) Similarly, incentives do not seem to alter the extent of communication as reported by teams (seven point Likert scale; mean (std. dev.) in *BGIncentive45*: 5.60 (1.28) versus 5.62 (1.39) in *BGControl*, Mann-Whitney test, p-value $= 0.58$). Finally, we observe a suggestively large yet not statistically significant difference in the likelihood that teams select the most skilled person for the logical reasoning task (84 percent in *BGIncentive45* versus 77 percent in *BGControl*, χ^2 test, p-value $= 0.40$).

Overall, the above analyses on potential mechanisms highlight two important aspects. First, our post-experimental survey questions from the framed field experiment indicate that incentives created an increased demand for leadership, and recent evidence from field experiments conducted in the same setting confirms that such an increased demand for leadership causes substantial performance improvements (see Englmaier et al., 2021). Second, our analyses on experts' expectations provides additional guidance on interesting avenues for future research in terms of better understanding how incentives may affect different effort dimensions in non-routine tasks. Our surveys identified which effort dimensions experts consider relatively more important and thus suggest which dimensions future research may focus on in more detail. Our laboratory experiment complements this approach by showing that incentive effects do not necessarily coincide with experts' expectations. Among the top three dimensions, we could only find suggestive evidence for one dimension (*skill-to-task matching*). While there remain many interesting avenues for further exploration of the precise relationship between incentives and

effort dimensions in non-routine tasks, taken together our results indicate that incentives likely improve team performance in non-routine tasks due an endogenous emergence of demand for leadership that goes hand in hand with changes in team organization (see Englmaier et al., 2021).

5 Discussion

Our results demonstrate that bonus incentives have sizable positive effects on team performance in both the natural and the framed field experiments. Following important work by Maniadis et al. (2014), we investigate how much our findings should update our beliefs that incentives truly increase performance in our task. To do so, we calculate Post-Study-Probabilities (PSPs) conditional on different priors. PSP $= (1-\beta)\pi/[(1-\beta)+\alpha(1-\pi)]$, where π denotes the probability of a given prior and $(1-\beta)$ the study's statistical power. Intuitively, the PSP reflects the posterior probability that our null hypothesis (no incentive effects) is false.

The results are displayed in Table 12, where the rows display increasing priors and the columns reflect different levels of power. Column 1 shows posteriors given statistical power of $(1-\beta) = 0.45$. This corresponds to the achieved power of our binary measures to complete the task within 45 or 60 minutes from our framed field experiment with the student sample. The posteriors indicate that even with moderate power, we should drastically update our beliefs upwards. Starting from priors as low as $\pi = 0.10$, which indicate a strong disbelief in any effect, the posteriors reflect equal probabilities of both outcomes (PSP$= 0.50$). Higher priors, including those favoring no effect, yield posteriors strongly siding with our result. Column 2 shows posteriors for a power of $(1-\beta) = 0.7$, which corresponds to our binary outcome variable on succeeding in 45 minutes for the natural field experiment. Column 3 reports posteriors for a power of $(1-\beta) = 0.95$, which we achieve for our binary outcome variable on succeeding in 60 minutes in the natural field experiment, as well as for t-tests on the remaining time in both the framed and the natural field experiment. Both columns show that even moderate to high disbelief converts into posteriors strongly favoring an effect to exist. In the survey described in Section 4.2, we also asked respondents if they believed incentives to influence performance in non-routine analytical team tasks. Over 80% believe that incentives have at least some positive effect, suggesting that assuming medium to high priors seems more

Table 12: Post-Study Probabilities

Achieved power for...	χ^2-tests on success dummy (45 & 60 mins) in framed field (1)	χ^2-tests on success dummy (45 mins) in nat. field (2)	χ^2-tests on success dummy (60 mins, nat. field) and t-tests on remaining time (field and framed field) (3)
	0.45	0.70	0.95
Prior probability	Posterior	Posterior	Posterior
0.05	0.32	0.42	0.50
0.10	0.50	0.61	0.68
0.20	0.69	0.78	0.83
0.35	0.83	0.88	0.91
0.50	0.90	0.93	0.95
0.75	0.96	0.98	0.98
0.90	0.99	0.99	0.99

Notes: This table reports Post-Study-Probabilites (Maniadis et al., 2014) for different combinations of prior probabilities and achieved power. The levels of power in columns 1 to 3 correspond to the achieved power in terms of statistical tests (t-tests and χ^2 tests) for our primary outcomes. We achieved a power of about 0.95 for t-tests on the remaining time in the natural and framed field experiment, as wells as for the χ^2-tests of whether the team received the bonus in the natural field experiment. Our achieved power for χ^2-tests of whether teams complete the task in 45 minutes amounts to 0.7 in the field experiment. In the framed field experiment, achieved power for the χ^2-tests of whether the team completes the task in 45 or 60 minutes respectively amounts to 0.45.

adequate. Overall, these results emphasize that one would need to very strongly believe in incentives being ineffective in these tasks to continue holding such a belief in the light of our findings.

We also deem it worth discussing that we observed the framing of incentives to be of minor importance. A loss frame did not generally outperform a gain frame but turned out to be statistically significantly larger only when considering the fraction of customer teams finished before 45 minutes (in some specifications). This result is in line with much of the literature, where significant framing effects have been observed in some environments (e.g. Fryer et al., 2012; Hossain and List, 2012; Muralidharan and Sundararaman, 2011), but not in others (DellaVigna and Pope, 2017).

What is driving the observed performance increase? With respect to hint taking, we have several reasons to believe that hints are not responsible for the observed differences in performance. 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 task and how many hints teams have taken before that step, we observe signif-

icantly better performance by teams facing incentives but no significant differences in hint taking (see Appendix A.8 and Table A.12).

An alternative possible explanation for how bonuses improve performance is that incentives 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 into performance. While we primarily designed our field experiments 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 incentives have a stronger effect on performance of teams with prior experience (see model (4) in Table A.5), 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. In line with recent evidence from field experiments that identify the causal effects of leadership in the same setting (Englmaier et al., 2021), these results render changes in team organization due to incentives a particularly important channel for why incentives improve team performance.

6 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. The experiment 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 sam-

ples are more than twice as likely to complete the task in 45 minutes under the incentive condition than under the control condition.

Our findings complement evidence from recent field experiments that have investigated the effects of bonus incentives in routine tasks (Friebel et al., 2017; Hossain and List, 2012). Although bonus incentives are simpler than optimal incentives generally suggested by standard theories, firms frequently use bonuses (Oyer, 2000). Apart from potential practical implementation advantages, recent theoretical contributions have considered additional motives for why simple binary payment schemes may be popular. On the one hand, theoretical models have been adjusted for important behavioral components such as fairness preferences (Fehr et al., 2007), overconfidence (Larkin and Leider, 2012), or loss aversion (Herweg et al., 2010). On the other hand, it has been argued that bonus incentives are particularly attractive for non-routine analytical tasks, as these often require searching for ideas in environments in which agents have an informational advantage and principals are unable to observe search behaviors. Ulbricht (2016) shows that if both asymmetries hold, search is almost surely inefficient whereas a menu of simple bonus contracts can still achieve the second-best solution. Hence, the choice of incentives needs to take behavioral aspects as well as the characteristics of the specific task performed into account.

To test for these behavioral aspects, we exploit a number of additional treatment variations in our natural field experiment. First, we implement the bonus incentives both in a gain and in a loss frame and find that framing team bonuses as a loss has at most a modest additional effect on performance, and only does so for a subset of our data. 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.[33] Third, we find that teams tend to be less likely to explore on their own when facing bonus incentives. However, this was mainly true 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,

[33]This does not rule out that salient performance goals may further increase team performance, as observed, e.g., in laboratory (Corgnet et al., 2015) and field experiments (Gosnell et al., 2020).

2013), by highlighting a potential relationship between the consequences of incentives for exploratory behavior and the intrinsic motivation to complete 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 increased demand for leadership within teams, and may thus change the way teams are organized.

Our study constitutes, to the best of our knowledge, the first systematic investigation into incentive effects in non-routine analytical and jointly solved team tasks. While we used a unique setting, which allowed for exogenous variation of incentives for a large number of teams and provided an objective measure of performance, we wish to discuss some potential caveats of our study. One important aspect is that real life escape games are primarily designed as a leisure activity (although they are sometimes also used for recruiting or as team-building events). Regular customers may thus be highly intrinsically motivated to work on the task, which may color their response to incentives. To test for this aspect explicitly, we replicated our main incentive treatments with student participants whom we hired and paid to perform the task. Similar to customers, these teams are more likely to complete the task and to do so faster with incentives.

Further, one could worry that our treatment effect may only arise when subjects are highly intrinsically motivated. This would be the case if the specific nature of the task raised the student's sample intrinsic motivation to the level of the customer sample. However, we observe a striking difference between the two samples with respect to hint taking, suggesting that both samples' intrinsic preference for completing the task differs. Finally, even if intrinsic interest in the task would be necessary for the treatment effect to arise, many real world work environments, particularly those featuring non-routine analytical team tasks, may be subject to self-selection of employees who acquire specific skills and apply for positions in which they can work on challenging but interesting tasks. As such, intrinsically motivated agents seem to be a feature of these jobs in general, rather than only a particular feature of our setting.

Our results raise interesting questions for future research. 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. A very interesting, but particularly challenging question that remains is to empirically find the optimal incentive mechanism for performance in non-

routine analytical team tasks. This requires varying different types of incentives (tournaments, bonuses, etc.) and their extent simultaneously, ideally on a set of non-routine tasks of different nature. While clearly beyond the scope of the current study, it is certainly a very interesting and relevant avenue for future research. Looking beyond the question of incentives, the setting of a real-life escape game may further 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*.

Azmat, G. and Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94(7):435–452.

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.

Barankay, I. (2010). Rankings and social tournaments: Evidence from a field experiment. Working Paper.

Barankay, I. (2012). Rank incentives evidence from a randomized workplace experiment. Working Paper.

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

Casner-Lotto, J. and Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century US workforce.* ERIC.

Charness, G. and Grieco, D. (2019). Creativity and incentives. *Journal of the European Economic Association*, 17(2):454–496.

Churchill, G. A., Ford, N. M., and Walker, O. C. (1993). *Sales Force Management: Planning, Implementation,and Control.* Irwin/McGraw-Hill,Homewood,Illinois.

Corgnet, B., Gómez-Miñambres, J., and Hernán-Gonzalez, R. (2015). Goal setting and monetary incentives: When large stakes are not enough. *Management Science*, 61(12):2926–2944.

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.

Deming, D. and Kahn, L. B. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, 36(S1):S337–S369.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, 132(4):1593–1640.

Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review*, 100(3):1238–60.

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., Grimm, S., Grothe, D., Schindler, D., and Schudy, S. (2021). The value of leadership: Evidence from a large-scale field experiment. Working Paper.

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.

Fehr, E., Klein, A., and Schmidt, K. M. (2007). Fairness and contract design. *Econometrica*, 75(1):121–154.

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. *IZA Discussion 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(1):489–521.

Gibbs, M., Neckermann, S., and Siemroth, C. (2017). A field experiment in motivating employee ideas. *Review of Economics and Statistics*, 99(4):577–590.

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.

Gosnell, G. K., List, J. A., and Metcalfe, R. D. (2020). The impact of management practices on employee productivity: A field experiment with airline captains. *Journal of Political Economy*, 128(4):1195–1233.

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

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. *JSAS Catalog of Selected Documents in Psychology*, 8:35.

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

Herweg, F., Müller, D., and Weinschenk, P. (2010). Binary payment schemes: Moral hazard and loss aversion. *American Economic Review*, 100(5):2451–77.

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.

i Vidal, J. B. and Nossol, M. (2011). Tournaments without prizes: Evidence from personnel records. *Management Science*, 57(10):1721–1736.

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

Jerald, C. D. (2009). Defining a 21st century education. *Center for Public education*, 16.

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.

Larkin, I. and Leider, S. (2012). Incentive schemes, sorting, and behavioral biases of employees: Experimental evidence. *American Economic Journal: Microeconomics*, 4(2):184–214.

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

Lazear, E. and Oyer, P. (2013). Personnel economics. In Gibbons, R. and Roberts, J., editors, *Handbook of Organizational Economics*, pages 479–519. Princeton University Press.

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.

List, J. A. (2003). Does market experience eliminate market anomalies? *The Quarterly Journal of Economics*, 118(1):41–71.

List, J. A. (2004a). The nature and extent of discrimination in the marketplace: Evidence from the field. *The Quarterly Journal of Economics*, 119(1):49–89.

List, J. A. (2004b). Neoclassical theory versus prospect theory: Evidence from the marketplace. *Econometrica*, 72(2):615–625.

List, J. A. (2006). The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of political Economy*, 114(1):1–37.

List, J. A., Shaikh, A. M., and Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 22(4):773–793.

Maniadis, Z., Tufano, F., and List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, 104(1):277–90.

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.

Moynahan, J. K. (1980). *Designing an effective sales compensation program.* Amacom, 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.

NACE (2015). Job outlook: National association of colleges and employers.

Oyer, P. (2000). A theory of sales quotas with limited liability and rent sharing. *Journal of labor Economics*, 18(3):405–426.

Pink, D. (2009). Dan Pink: The puzzle of motivation. `https://www.ted.com/talks/danpinkonmotivation`. Accessed: 2018-03-05.

Pink, D. H. (2011). *Drive: The surprising truth about what motivates us*. Riverhead Books, New York.

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

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.

Ulbricht, R. (2016). Optimal delegated search with adverse selection and moral hazard. *Theoretical Economics*, 11(1):253–278.

A Supplementary Appendix

A.1 Room Fixed Effects for the Natural and Framed Field Experiment

Table A.1: Main treatments probit and GLM regressions including room fixed effects

	Field experiment (1)-(2)		Framed field experiment (3)-(4)	
	Probit (ME)	GLM	Probit (ME)	GLM
	(1)	(2)	(3)	(4)
Bonus45	0.150***	0.266**	0.076**	0.655***
	(0.041)	(0.113)	(0.036)	(0.215)
Constant		3.706***		3.896***
		(0.488)		(0.834)
Fraction of control teams completing 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

Notes: The table shows average marginal effects from probit regressions of whether a team completed 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 (4), A.7 (4), 7 (4), and A.9 (4), but include in addition Room Fixed Effects. Robust standard errors clustered at the day (field experiment) and session (framed field experiment) level reported in parentheses, and *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

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 escape 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 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.4 Multiple Hypotheses Testing (adjusted p-values)

A.4.1 Field Experiment

Table A.2 presents p-values adjusted for multiple hypotheses testing according to Theorem 3.1 in List et al. (2019), by simultaneously testing for differences in multiple outcomes and treatments (where appropriate). For the pooled treatment effect (*Bonus45* vs. *Control*), we correct for multiple outcomes. For the effects of *Gain45* and *Loss45*, we correct for multiple outcomes and treatments and perform all pairwise comparisons simultaneously. The pooled treatment effect is still significant at the 1-percent level for all four outcome variables. Both *Gain45* and *Loss45* significantly increase the fraction completing the task within 45 minutes and significantly reduce unconditional and conditional remaining times. Solely the fraction of teams finishing the task within 60 minutes in *Gain45* (vs. *Control*, p-value$= 0.1443$) and *Loss45* (vs. *Control*, p-value$= 0.1050$) fails to

Table A.2: Field experiment - MHT adjusted p-values according to List et al. (2019) (referring to Table 4)

Outcome	*Control* vs. *Bonus45*	*C.* vs. *Gain45*	*C.* vs. *Loss45*	*Gain45* vs. *Loss45*
Fraction completing in 45 min	0.0003	0.0073	0.0003	0.7773
Fraction completing in 60 min	0.0083	0.1050	0.1443	0.8523
Mean remaining time (in sec)	0.0003	0.0003	0.0080	0.8367
Mean r. time (in sec) if completed	0.0010	0.0173	0.0523	0.8343

Notes: This table shows p-values adjusted for multiple hypotheses testing according to (List et al., 2019) for comparisons of *Control* vs. the pooled bonus incentive treatment (*Bonus45*) (corrected for multiple outcomes), and *Control* vs. *Gain45*, *Control* vs. *Loss45*, and *Gain45* vs *Loss45* adjusted for multiple outcomes and treatments testing for all pairwise comparisons.

Table A.3: Field experiment - MHT adjusted p-values according to List et al. (2019)

Outcome	*Control* vs. *Bonus45*	*C.* vs. *Bonus60*	*C.* vs. *Reference Point*
Fraction completing in 45 min	0.0003	0.2030	0.8943
Fraction completing in 60 min	0.0543	0.2203	0.9080
Mean remaining time (in sec)	0.0003	0.3570	0.9850
Mean r. time (in sec) if completed	0.0003	0.8717	0.9260

Notes: This table shows p-values adjusted for multiple hypotheses testing according to (List et al., 2019) for comparisons of *Control* vs. *Bonus45*, *Control* vs. *Bonus60*, and *Control* vs *Reference Point* adjusted for multiple outcomes and treatments.

differ significantly at the ten percent level when performing twelve tests simultaneously. Outcomes in *Gain45* and *Loss45* treatments do not differ.

Table A.3 relates to Table 5 and presents adjusted p-values by simultaneously testing for differences in multiple outcomes and treatments (*Bonus45*, *Bonus60*, and *Reference Point* to *Control*). Our main treatment *Bonus45* is still significant at conventional levels. The increase in the fraction of teams finishing the task (in 45 or 60 minutes) in *Bonus60* and the reduction in remaining times is too small to reach significance at conventional levels when adjusting p-values conservatively for twelve simultaneous tests. However, even these adjusted p-values are substantially smaller than the p-values for the *Reference Point* treatment, which has essentially no effect on the four outcome variables. Hence, our conclusion that we do not observe any performance effects solely due to introducing reference points remains.

Table A.4: Framed Field experiment - MHT adjusted p-values according to List et al. (2019) (referring to Table 6)

Outcome	*Control* vs. *Bonus45*	*C.* vs. *Gain45*	*C.* vs. *Loss45*	*Gain45* vs. *Loss45*
Fraction completing in 45 min	0.0830	0.2163	0.6720	0.6687
Fraction completing in 60 min	0.0520	0.5837	0.0883	0.443
Mean remaining time (in sec)	0.0023	0.0807	0.0107	0.8353
Mean r. time (in sec) if completed	0.0320	0.0547	0.2123	0.6913

Notes: This table shows p-values adjusted for multiple hypotheses testing according to (List et al., 2019) for comparisons of *Control* vs. the pooled bonus incentive treatment (*Bonus45*) (corrected for multiple outcomes), and *Control* vs. *Gain45*, *Control* vs. *Loss45*, and *Gain45* vs *Loss45* adjusted for multiple outcomes and treatments testing for all pairwise comparisons.

A.4.2 Framed Field Experiment

Table A.4 refers to Table 6 and shows p-values adjusted for multiple hypotheses testing according to Theorem 3.1 in List et al. (2019), by simultaneously testing for differences in multiple outcomes and treatments (where appropriate) for the framed field experiment. After adjusting p-values for testing on multiple outcomes, the pooled treatment effect is still significant at conventional levels for all four outcome variables. Further, the remaining times significantly differ between *Gain45* and *Control* and *Loss45* and *Control* when correcting for testing on multiple outcomes and all pairwise comparisons simultaneously.

A.5 Additional Analyses for the Field Experiment

A.5.1 Bonus Incentives and Team Characteristics

Table A.5 shows the results from linear probability models estimating a dummy for whether teams complete 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 in model (4) turns out to be significant (at the five percent level) and positive, while at the same time the treatment dummy is still statistically significant and large in magnitude. Hence, the positive incentive effect is robust and slightly larger for teams with experience.

Table A.5: Linear probability model: Completed in less than 45 minutes

	OLS: Completed in less than 45 minutes					
	(1)	(2)	(3)	(4)	(5)	(6)
Bonus45	0.172***	0.200***	0.023	0.120**	0.130**	0.169***
	(0.050)	(0.071)	(0.122)	(0.057)	(0.056)	(0.047)
Share males	0.102*	0.130**	0.102*	0.100*	0.105*	0.103*
	(0.055)	(0.048)	(0.055)	(0.054)	(0.056)	(0.058)
Group size	0.056***	0.056***	0.042**	0.057***	0.055***	0.056***
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
Experience	0.125***	0.126***	0.126***	0.058*	0.124***	0.125***
	(0.031)	(0.031)	(0.032)	(0.032)	(0.031)	(0.031)
Private	0.040	0.039	0.039	0.036	−0.001	0.039
	(0.041)	(0.042)	(0.042)	(0.041)	(0.049)	(0.041)
English-speaking	−0.115*	−0.117*	−0.113*	−0.114*	−0.117*	−0.129***
	(0.060)	(0.062)	(0.062)	(0.060)	(0.059)	(0.044)
Bonus45 ...						
... × Share males		−0.055				
		(0.128)				
... × Group size			0.031			
			(0.025)			
... × Experience				0.132**		
				(0.051)		
... × Private					0.077	
					(0.056)	
... × English speaking						0.027
						(0.139)
Constant	−0.177	−0.192	−0.109	−0.179	−0.163	−0.172
	(0.132)	(0.151)	(0.142)	(0.132)	(0.133)	(0.138)
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

Notes: 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 clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.5.2 Probability of Completing the Task in 45 Minutes (Field Experiment)

Table A.6 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. This confirms our previous findings.

Table A.6: Main treatments probit regressions: Excluding weeks with no variation in the outcome variable

	Probit (ME): Completed in less than 45 minutes				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.150***	0.151***	0.183***	0.163***	
	(0.026)	(0.024)	(0.027)	(0.045)	
Gain45					0.134***
					(0.040)
Loss45					0.188***
					(0.050)
Fraction of control teams completing 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

Notes: The table reports average marginal effects from probit regressions of whether a team completed the task within 45 minutes on our treatment indicators (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 clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.5.3 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 cluster-robust standard errors (Table A.7). 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. Finally, in Column (5) we include

Table A.7: GLM regressions: Remaining time

	GLM: Remaining time in seconds				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.432***	0.447***	0.406***	0.257**	
	(0.088)	(0.096)	(0.094)	(0.116)	
Gain45					0.259**
					(0.108)
Loss45					0.256*
					(0.136)
Constant	5.842***	4.041***	4.251***	3.803***	3.803***
	(0.082)	(0.393)	(0.359)	(0.403)	(0.403)
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

Notes: 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 clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

two treatment dummies to test whether gain or loss frames affect performance differently. Both coefficients are of similar size and we cannot reject the equality of the coefficients for the *Loss45* and *Gain45* treatments (Wald test, p-value = 0.98).

Analogously to the probit regressions reported in Table 5, we also run GLM specifications with the remaining time as the dependent variable (Table A.8) 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.8: GLM regressions: Remaining time (all treatments)

	GLM: Remaining time in seconds			
	(1)	(2)	(3)	(4)
Bonus45	0.432***	0.436***	0.376***	0.244**
	(0.088)	(0.093)	(0.092)	(0.102)
Bonus60	0.233*	0.267**	0.392***	0.449***
	(0.131)	(0.114)	(0.126)	(0.134)
Reference Point	0.002	−0.001	0.102	0.131
	(0.106)	(0.108)	(0.114)	(0.086)
Constant	5.842***	4.044***	4.225***	3.713***
	(0.081)	(0.317)	(0.310)	(0.329)
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

Notes: 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 clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.6 Additional Analyses for the Framed Field Experiment

A.6.1 Regression Analysis for Remaining Time as Dependent Variable (Framed Field Experiment)

Table A.9 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 remain roughly unchanged with the addition of controls and fixed effects. Column (5) shows the regression on the non-pooled framing treatments. The coefficients for both frames are highly significant but equality of coefficients of *Gain45* and *Loss45* cannot be rejected (p-value $= 0.88$).

Table A.9: GLM regressions: Remaining time (student sample)

	GLM: Remaining time in seconds				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.658***	0.673***	0.664***	0.661***	
	(0.216)	(0.217)	(0.210)	(0.213)	
Gain45					0.676***
					(0.238)
Loss45					0.647***
					(0.226)
Constant	5.135***	3.816***	4.039***	3.684***	3.690***
	(0.195)	(0.678)	(0.723)	(0.894)	(0.889)
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

Notes: 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 7. Robust standard errors clustered at the session level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.6.2 Probability of Completing the Task in 45 Minutes (Framed Field Experiment)

Table A.10 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. This confirms our previous findings.

Table A.10: Main treatments probit regressions: Excluding weeks with no variation in the outcome variable (student sample)

	Probit (ME): Completed in less than 45 minutes				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.107*	0.097*	0.104**	0.111**	
	(0.055)	(0.054)	(0.052)	(0.051)	
Gain45					0.142**
					(0.057)
Loss45					0.072
					(0.055)
Fraction of control teams completing 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

Notes: The table reports average average marginal effects from probit regressions of whether a team completed the task within 45 minutes on our treatment indicators (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 clustered at the session level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

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

Table A.11: Ordered probit regressions: Number of hints requested

	Ordered probit: Number of hints requested							
	Field experiment (1)-(4)				Framed field experiment (5)-(8)			
	within 60 min		within 45 min		within 60 min		within 45 min	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bonus45	0.116	0.086	0.341**	0.190	0.401***	0.395***	0.878***	0.933***
	(0.123)	(0.148)	(0.133)	(0.129)	(0.151)	(0.148)	(0.144)	(0.147)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Staff FE	No	Yes	No	Yes	No	Yes	No	Yes
Week FE	No	Yes	No	Yes	No	Yes	No	Yes
Observations	487	487	487	487	268	268	268	268

Notes: Coefficients from an ordered probit model of the number of hints requested within 60 minutes or 45 minutes regressed on our treatment indicator *Bonus45* (pooled). Controls and fixed effects (FE) identical to previous tables. Robust standard errors clustered at the day (field experiment) and at the session (framed field experiment) level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.8 Hint Taking at a Specific Step in the Task

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 they are likely to reach a difficult quest faster than slower teams. That is, if incentives make (worse) teams faster, these teams may also mechanically take more hints and this effect 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. This allows us to control the number of quests solved and to relate fixed progress in the task to hints taken. 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 escape 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 condition outperform control teams also early in the task. In Table A.12 we

Table A.12: Ordered probit regressions: Number of hints taken when entering last room (field experiment)

	Ordered probit: Number of hints taken				
	(1)	(2)	(3)	(4)	(5)
Bonus45	−0.018	0.012	0.113	0.050	0.134
	(0.115)	(0.113)	(0.084)	(0.110)	(0.137)
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

Notes: 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 clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

report results from ordered probit models to study whether teams facing incentives take more 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 the intermediate step 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.13).

Table A.13: Ordered probit regressions: Number of hints taken when entering last room (framed field experiment)

	Ordered probit: Number of hints taken				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.244**	0.235*	0.285**	0.306***	0.361**
	(0.122)	(0.123)	(0.119)	(0.117)	(0.154)
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

Notes: 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 clustered at the session level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.9 Hint Taking and Risk Aversion

One might be concerned that original solutions may be perceived as riskier, in particular when incentives are at play. In order to reduce exposure to such risks, participants from the student sample (who may be differently risk-averse to customers) simply request more hints under incentives, thus mechanically inducing the difference in requested hints across treatment condition. Fortunately, the data from our framed field experiment allows us to test whether heterogeneity in the willingness to take risks is decisive for hint taking, and whether incentives interact with the willingness to take risks. Using our measure for risk taking in general (Dohmen et al., 2010), we regress the number of hints taken (within 60 and 45 minutes) on the incentive condition, whether the teams' propensity to take risk lies above or below the median and the interaction between these two explanatory variables. Table A.14 shows that both below median risk-taking and the interaction term do not significantly affect hint-taking behavior. Models (2) and (4) show the same results but include additional controls as well as host and week fixed effects. All columns show that risk preferences appear to play a minor role in terms of magnitude and significance (compared to the treatment) and do not interact significantly with incentives. Hence, we deem it unlikely that greater risk aversion coupled with bonus incentives leads to fewer original solutions in our setting.

Table A.14: OLS regressions: Number of hints requested

	Dependent variable: Number of hints requested within			
	60 mins		45 mins	
	(1)	(2)	(3)	(4)
Bonus45	0.394*	0.356*	0.811***	0.815***
	(0.200)	(0.186)	(0.168)	(0.161)
Below median willingness to take risks	0.009	0.024	0.099	0.192
	(0.245)	(0.231)	(0.195)	(0.206)
Bonus45 x	-0.046	-0.027	0.057	-0.029
below median willingness to take risks	(0.283)	(0.274)	(0.248)	(0.254)
Constant	3.735***	4.713***	2.286***	3.007***
	(0.174)	(0.736)	(0.132)	(0.668)
Control Variables	No	Yes	No	Yes
Staff Fixed Effects	No	Yes	No	Yes
Week Fixed Effects	No	Yes	No	Yes
Room Fixed Effects	No	Yes	No	Yes
Observations	268	268	268	268
R-squared	0.030	0.175	0.139	0.292

Notes: Coefficients from OLS regressions of the number of hints requested in the framed field experiment within 60 minutes or 45 minutes regressed on our treatment indicator *Bonus45* (pooled), whether the team' propensity to take risk in general lies above or below the median, and the interaction of those variables. Controls and fixed effects (FE) identical to previous tables. Robust standard errors clustered at the session level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.10 Additional laboratory experiment: Description

Our laboratory experiment is based on a board game version of a real life escape game. The board game resembles similar features as our field setting but allows us to alter some sub-tasks to explicitly study the causal effects of team incentives on the three effort dimensions our survey respondents considered most important: First, we test if incentives causally affected whether teams assign the most skilled team member to a specific subtask (*skill-to-task matching*). Second, we investigate the causal effect of incentives on the likelihood of team members sharing relevant information (*information sharing*) to facilitate task completion. Third, we study the causal effect of incentives on *communication*.

The non-routine team task is framed as a secret mission, in which participants need to gain access to the palace of the leader of a fictitious country (part I), find some secret information in the palace (part II), and escape (part III), all within 60 minutes. Each part contained several sub-parts (e.g. part I.2 denotes sub-part 2 of part I). As participants arrive at the laboratory, they are randomized into teams of three and each team is guided a separate room to perform the task (with treatments being randomized across these rooms as well). In each room, one experimenter welcomes the participants and explains the general procedures, before each participant undergoes a cognitive skill test (Raven's progressive matrices) on a computer tablet at a separated workstation. After completion of the test, each participant receive her own test score as private information but no participant is informed about their team members' performance in the test. Then, all three participants are guided to stand around a large table in the middle of the room, to perform the boardgame escape challenge.

For the boardgame challenge, participants were guided by a tablet computer placed in the middle of the table. The tablet displays the time left to solve the escape challenge and serves to electronically store task solutions entered by the team. Further, the tablet displays hints to help teams make progress at pre-specified times (that is, all teams received the exact same hints at the exact same time, a feature adapted from the board game our team challenge is based on). To take notes, each participant receives a pen and a paper, and each team member is equipped with an identical decoding sheet. Further, each team member receives an envelope with a text containing information about the layout of the leader's office in the palace. This text mostly contained useless but entertaining information, but also, and different for each team member, some information that could

help to find the solution to part II.2. Participants were explicitly told that they are not allowed to share this information at that stage but were not explicitly informed that this information could help to solve part II.2 much faster, when combined.

After participants indicated they are ready to commence with the experiment, a 60 minute clock was started on the tablet computer and the team received an envelope containing the materials for part I.1. These materials included a name tag with an empty field at the bottom titled 'personal code', an invitation letter to the palace opening containing the information to 'bring your personal code', another decoding sheet displaying a matrix of numbers, several keys, and a white paper strip with small dots and stripes on both sides. At this stage, the tablet computer asked participants to enter their personal code, which could be found by combining the dots and stripes shown on both sides of the paper strip. The resulting pattern could then be decoded (using the decoding sheet distributed initially) to obtain the personal code.[34]

After completion of this part, subjects advanced to part I.2 and subsequently to part II.1. We designed parts I.2 and II.1 to be similar, yet challenging to subjects.

The materials for part I.2 consisted of 5 different flags, an invitation card reminding subjects not to speak (if communication was prohibited in part I.2), a text of the country's national anthem, and a note from the country's leader, saying that the combination of the country's flag and the personal code would yield the solution to part I.2. To arrive at the solution, participants had to study the anthem's text to identify the correct flag.[35] They could then use the decoding sheet from part I.1 to identify the correct four-digit number needed to solve the quests in part I.2. Using the keys handed out in part I.1 (which bore single digit numbers), subjects needed to select the four keys (in the right order) on the tablet computer to end part I.[36] After they managed to do so, the experimenter distributed materials for part II.1.

In part II.1, participants received information cards for five different fictitious enemy countries (with a map of each country and some basic info such as GDP, strength of armed forces, and other information), a solution sheet containing a matrix that would yield two of the four correct keys to terminate part II, and a speech by the leader describing the country he considered to be the greatest enemy (containing a reminder not to speak should verbal communication be prohibited in part II.1). Selecting the greatest enemy

[34] If participants failed to enter the correct code, 3 minutes were substracted from their available time.
[35] Each time participants chose the wrong flag, 3 minutes were substracted from their available time.
[36] If participants failed to enter the correct key code, 1 minute was substracted from their available time.

country could be achieved by combining clues from the speech with the information on the country information cards and then making use of the matrix on the solution sheet.[37]

Verbal communication was randomly prohibited in either part I.2 or part II.1, and this was announced only at the beginning of the respective part. The communication ban was implemented by the experimenter under the threat of exclusion and, after the respective subpart was solved, the experimenter also immediately announced that the team could again communicate. In half of all sessions, the contents of part II.1 and part I.2 were exchanged to avoid order effects. This exogenous variation of the availability of verbal communication was introduced to allow for an analyses of the effects of incentives on performance through communication in a diff-in-diff analyses.[38]

In part II.2, subjects could share the information distributed before the start of the experiment. Importantly, the information provided was sufficient, but not necessary to arrive at the correct solution. Alternatively, subjects could decide to not share their information and use the materials provided to work on the part's solution. By comparing the differences in how much information was shared across treatments with and without incentives, this subpart allows us to determine the causal effect of incentives on *information sharing*.

The materials for part II.2 were a picture of the leader's office, as well as instructions to 'count the golden eagles' displayed there, as well as a sheet translating Roman into Arabic numerals. Participants could simply search for all golden eagles in the picture, but they could also arrive at the solution by sharing the information they received prior to the experiment. Two of the three participants received information about the number of golden eagles in certain parts of the room at the beginning of the experiment, which combined yielded the total number. This number, translated into Roman numerals yielded the last two keys, as all keys (in addition to single digit Arabic numbers) also each bear a Roman numeral. Entering all four keys on the tablet computer ended part II.[39]

For part III, subjects were explicitly asked to select a team member for an individual task requiring logical reasoning. They were not reminded of their cognitive skill

[37] Each time participants chose the wrong enemy country, 3 minutes were substracted from their available time.

[38] As we do not find that incentives significantly affect the extend of communication reported by our participants, we refrain from including such a diff-in-diff analyses in the main text. Further, we do not find any indication that incentives significantly affect the difference in times needed to solve the sub-tasks in part II.1 and part I.2 with (vs. without) communication (p-value$= 0.30$, Mann–Whitney test).

[39] Each time participants entered a wrong key code, 1 minute was substracted from their available time.

test results obtained before the experiment and not made aware of a possible correlation between ability to perform in the individual task of part III and this test. They could, however, themselves take the initiative and discuss the results if they so wished. By comparing whether teams are more likely to choose the team member with the highest score with rather than without incentives conditions, we can estimate the causal effect of incentives on *skills-to-task matching*. After the team decided for a member, this member was guided to a secluded desk, where she received th respective materials and instructions. The individual task required to sort 8 picture cards (with pictures on both sides) into a 2×4 matrix based on a number of logical statements accompanying the instructions (e.g. 'the green flower pot can never be next to the green portrait'). By combining all statements, only one possible solution for arranging the picture cards remained.[40] Meanwhile, the remaining two group members worked on a variety of diverse tasks. They needed to detect a pattern in a sequence of numbers and continue the sequence, find an object hidden in a stereoscopic image, arrange keys in a specific fashion so they form the shape of a number, and use a key to follow a drawn path on a paper slip to unveil some letters. The solutions to these four tasks yielded the four keys to end part III and thus the game, while the solution to the individual task done by the third team member yielded the order in which the keys had to be entered.[41]

After participants entered the correct four keys (or if the 60 minutes expired, whichever occured first), the task ended and participants filled in a sort survey. The survey included questions related to why a specific person was chosen for the individual task in part III, questions on leadership as in the framed field experiment, a rating of the statements regarding all ten dimensions used in the expert and ESA survey, as well as general demographics such as age and gender and experience with escape room (board) games. If participants were assigned to a bonus condition and managed to (did not manage to) complete the task within 45 minutes, they received (kept) the bonus payment in *BGGain45* (in *BGLoss45*). Otherwise they did not receive the bonus (or handed it back in *BGLoss45*). All participants also received the participation fee and were subsequently dismissed from the laboratory.

Our power calculations for the additional laboratory experiment were based on our findings in the framed field experiment (student sample) and on assumptions about the

[40] Each time the participant entered a wrong solution, 1 minute was substracted from the available time.
[41] Each time participants entered a wrong key code, 1 minute was substracted from their available time.

data generating process and performances in the respective sub-tasks of the additional laboratory experiment. A sample of 120 groups (with 40 groups in *BGGain45*, 40 in *BGLoss45*, and 40 in *BGControl*) allows us to identify pooled incentive effect sizes of about 0.547 standard deviations in two-sample t-tests with statistical power of 80 percent at the five-percent significance level. That is, if we observe similar finishing times and variances as in the framed field experiments, we can identify effects of incentives (pooled) on the remaining time that are larger than 3 minutes an 13 seconds. As in our framed field experiment, power is expected to be lower for binary outcomes such as finishing within 60 or 45 minutes. Using a χ^2-test, we can identify effect sizes larger than 17 to 27 percentage points, depending on the fraction of subjects finishing the task in *BGControl* within 45 or 60 minutes.

Following these calculations, we recruited in total 381 participants to form 127 teams consisting of three members each. Due to technical problems with the experimental software, we need to discard three observations. In these sessions, subjects were not acoustically made aware of a hint being displayed, distorting their progress in the game relative to other participants. We remove another five sessions by one particular research assistant, as they did not administer the treatment correctly in at least one session and were the only research assistant (out of ten) to receive participants' complaints about not having properly delivered the instructions. This leaves us with 119 observations. Akin to the framed field experiment, we assigned roughly two thirds of teams to the incentive treatment (36 to *BGGain45*, 37 to *BGLoss45*) and roughly one third to *BGControl* (46). To avoid time trends in the data affecting our results, we ran three sessions concurrently whenever possible, to have each treatment present at any same time and day. Due to no-shows of participants, some slots featured fewer sessions.

The main aim of the additional laboratory experiment was to study whether incentives causally affect the three effort dimensions considered as most important by our survey respondents: *Skill-to-task matching*, *Information sharing* and *Communication*. To do so, we discuss in the main text whether bonus incentives alter the quality of *skill-to-task matching* (i.e. the likelihood of selecting the person with the highest cognitive test score in part III). Similarly, we study whether incentives affect the number of team members sharing information (which is received at the beginning of the experiment) in part II.2 (the "counting eagles" sub-task), and whether team members' report different levels of communication in the incentive condition (team members were individually asked at

the end of the experiment to what extend they agree with the statement "We communicated a lot" on a seven-point Likert scale ranging from "fully disagree" to "fully agree"). As we do not observe any significant treatment effects on these outcome variables, we refrain from presenting additional analyses planned to be run, had we observed stronger treatment effects (see also our pre-registration for the additional laboratory experiment).