The effect of incentives in non-routine analytical team tasks

Englmaier, Florian; Grimm, Stefan; Grothe, Dominik; Schindler, David; Schudy, Simeon

Published in:
Journal of Political Economy

Publication date:
2023

Document Version
Early version, also known as pre-print

CESifo WORKING PAPERS

6903
2018

Original Version: February 2018
This Version: April 2023

The Effect of Incentives in Non-Routine Analytical Team Tasks

Florian Englmaier, Stefan Grimm, Dominik Grothe, David Schindler, Simeon Schudy

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 6903*

The Effect of Incentives in Non-Routine Analytical Team Tasks

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 series of field experiments with more than 5,000 participants, we investigate how incentives alter behavior in teams working on such a task. We document a positive effect of bonus incentives on performance, even among strongly intrinsically motivated teams. Bonuses also transform team organization as they enhance teams' demand for leadership. Exogenously increasing the demand for leadership establishes a causal link that explains a large part of the observed bonus-induced performance improvements.

JEL-Codes: C920, C930, J330, D030, M520.

Keywords: team work, bonus, incentives, leadership, non-routine, exploration.

Florian Englmaier
University of Munich
Department of Economics & Organizations Research Group (ORG)
Germany – 80539 Munich
florian.englmaier@econ.lmu.de

Stefan Grimm
University of Munich
Department of Economics
Germany – 80539 Munich
stefan.grimm@econ.lmu.de

Dominik Grothe
University of Munich
Department of Economics
Germany – 80539 Munich
dominik.grothe@econ.lmu.de

David Schindler
Tilburg University
Department of Economics
The Netherlands – 5000 LE Tilburg
d.schindler@tilburguniversity.edu

Simeon Schudy
University of Munich
Department of Economics
Germany – 80539 Munich
simeon.schudy@econ.lmu.de

April 15, 2023

We thank Steffen Altmann, John Antonakis, Oriana Bandiera, Iwan Barankay, Erlend Berg, Jordi Blanes i Vidal, Laura Boudreau, Alexander Cappelen, Lea Cassar, Eszter Czibor, David Cooper, Robert Dur, Florian Ederer, Constança Esteves-Sorenson, Anastasia Danilov, Armin Falk, Urs Fischbacher,Wouter Dessein, Guido Friebel, Svenja Friess, Uri Gneezy, Holger Herz, David Huffman, Lorenz Götte, Simon Jäger, Rajshri Jayaraman, Steven Levitt, Botond Kőszegi, Michael Kosfeld, Andreas Leibbrandt, Stephen Leider, Rocco Macchiavello, Clarissa Mang, 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, Julian Angermaier, Christian Boxhammer, Thomas Calcagno, Florian Dendorfer, Silvia Fernandez Castro, Katharina Hartinger, Michael Hofmann, Simon Klein, Yutaka Makabe, Giuseppe Musillo, Timm Opitz, Julia Rose, Regina Seibel, Till Wickert, Nicolas Wuthenow, Aloysius Widmann for excellent research assistance. David Schindler gratefully acknowledges funding by the Dutch Research Council (NWO) under the Talent Scheme (file number VI.Veni.211E.002). Stefan Grimm acknowledges funding by the German Research Foundation (DFG) through GRK 1928. Financial support by the DFG through CRC TRR 190 (Project Number 280092119) 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).

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 the latter category 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 performance in such 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.[1]

[1]This study focuses on performance-related bonus payments which firms may use as part of their annual incentive plans. The 2021 CAP-WorldatWork Incentive Pay Practices Survey (`https://worldatwork.org/resources/research/incentive-pay-practices`) indicates that both short- and long-term incentives are prevalent among a variety of companies from different sectors (>90 percent of which use short- and >50 percent use long-term incentives) with on average, 76 percent of firms using annual incentive plans. However, the use of different annual incentive pay components varied substantially across firms and levels, rendering the question whether bonus incentives work in non-routine tasks crucial from a practitioners perspective. For a more general discussion on the use of performance-related bonus payments as part of compensation in firms see also Prendergast (1999), Lazear (2000), Oyer (2000) Lazear and Oyer (2013), Moynahan (1980), and Churchill et al. (1993). For theoretical motivations to use simple binary

There are several reasons why the efficacy of incentives may substantially differ in non-routine analytical team tasks. First, they are often performed by people who enjoy their challenging nature and are intrinsically motivated (see, e.g., Autor and Handel, 2013; Delfgaauw and Dur, 2010; Friebel and Giannetti, 2009).[2] In turn, extrinsic incentives could negatively affect team performance by crowding out workers' intrinsic motivation (e.g., Deci et al., 1999; Eckartz et al., 2012; Gerhart and Fang, 2015; Hennessey and Amabile, 2010). Bénabou and Tirole (2003) provide a theoretical framework formalizing arguments for crowding out based on the idea that incentives may alter workers' perception of the task or their own ability. For example, workers may infer from the existence of incentives that the task is less enjoyable than expected, or that incentives are likely implemented for less able or less intrinsically motivated workers.[3] Further, as non-routine tasks are generally multidimensional, incentives may lead to crowding out due to a substitution of effort (Holmstrom and Milgrom, 1991). As these tasks require information acquisition, information recombination, and creative thinking, there is thus room for performance incentives to discourage activities not included in the relevant performance measure, such as the autonomous exploration of new and original approaches (e.g., Amabile, 1996; Azoulay et al., 2011; Ederer and Manso, 2013; McCullers, 1978; McGraw, 1978).

Second, the efficacy of incentives may differ, as output could be a noisier function of effort than in routine tasks. In particular, optimal team production in non-routine tasks likely requires more coordination of individual efforts than in routine team tasks, potentially reducing the efficacy of any incentives that do not specifically stimulate such coordination. In a similar spirit, incentives may be less effective in non-routine tasks as workers may possess less knowledge about the production function, or because these tasks are typically found in fields for which employees may already have large incentives to perform well (due to intrinsic motivation, status, recognition, or career concerns). The

payment schemes, see, e.g., Fehr et al. (2007), Larkin and Leider (2012), Herweg et al. (2010), or Ulbricht (2016).

[2]Intrinsic motivation may stem from direct task utility (and thus reflect lower levels or lower marginal effort costs), or from benefits beyond the production outcome such as additional utility due to self- or social-signaling motives (Bénabou and Tirole, 2003, 2006), or greater goals that are attached to the activity (such as job mission, see, e.g., Cassar, 2019). We do not consider greater goals or job missions necessary ingredients in all non-routine team tasks. Still we believe that, both, direct task utility and benefits beyond the production outcome are often relevant in non-routine analytical team tasks. Even without greater goals, their challenging nature renders these tasks interesting, and, by performing well, agents can signal their ability (to themselves and others).

[3]As such, incentive effects may also interact with whether the task is perceived as interesting (Takahashi et al., 2016).

variety of reasons for why incentives may work differently in non-routine analytical tasks is mirrored in substantial heterogeneity in experts' expectations about the efficacy of incentives, calling for clean empirical evidence of how incentives alter behavior in teams jointly performing non-routine analytical tasks.[4]

This study exploits a unique field setting to measure the effects of bonus incentives for behavior in teams jointly performing a non-routine analytical task. We study the performance of teams in a real-life escape game in which they 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). The 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 permits 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. Further, our unique field setting is particularly flexible, all of which allows us to substantially advance the literature on incentives for jointly solved non-routine tasks.

To identify the causal effects of incentives on behavior, we first conducted a series of field experiments with strongly intrinsically motivated teams (customers of our cooperation partner) who were unaware of taking part in an experiment.[5] We implemented a between-subjects design, in which teams were randomly assigned to either a treatment or a control condition. For the main treatment, we offered a team bonus if the team com-

[4]For instance, we document that the range of HR experts' predictions of incentive efficacy varies strongly across experts. While the median HR expert expects 40 out of 100 teams to improve when facing incentives, 20% of them believe that between 0-20 teams will improve, while another 20% believe that 60-100 teams will improve (see Appendix Figure A.3 for the full distribution).

[5]Harrison and List (2004) classify this approach as a "natural field experiment". The study was approved by the Department of Economics' IRB at LMU Munich (Project 2015-11) and excluded customer teams with minors. In the general booking process, customers also gave written consent that data obtained at ETR could be shared with third parties for research purposes.

pleted 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. We find that bonus incentives significantly and substantially increase performance. 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 non-routine tasks feature an important noisy component in how effort translates into performance, bonus incentives did not only induce a local effect around the threshold for receiving the bonus, but instead improved performance over a significant part of the distribution of finishing times.

We then leverage the advantages of our setting and study in depth the most important aspects through which bonuses alter behavior in teams. To investigate the role of potential crowding out of intrinsic motivation, we use a three-pronged approach.

First, Bénabou and Tirole (2003) argue that incentives may alter workers' perceptions and thereby crowd out their intrinsic motivation to exert effort and perform well. Indeed, it seems plausible that bonus incentives can serve as negative signals about the task or a worker's type in our setting. Still, the results from our main treatment do not indicate substantial crowding out among strongly intrinsically motivated teams. However, our main treatment combined the bonus payment with a rather ambitious performance threshold (45 minutes), which could have been interpreted as a positive signal about workers' ability. Further, the ambitious performance threshold itself could have caused performance improvements (independently of the bonus incentive). To test for such countervailing effects, we implemented two additional treatment conditions. The first combined the bonus with a less ambitious performance threshold (60 minutes) and thus provides additional room for crowding out due to incentives. The second condition provided the ambitious (45 minutes) threshold as a reference point signaling excellent performance but no monetary reward. Results from these treatments reveal that the observed performance improvements clearly result from the monetary reward provided and do not depend on which reference point they were combined with.[6] Hence, it is unlikely

[6]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 work referring to workers' awareness of relative performance (see e.g., Blanes i Vidal and Nossol, 2011; Azmat and Iriberri, 2010; Barankay, 2010, 2012). Our finding, however, 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).

that the existence of the bonus incentive strongly crowded out teams' intrinsic motivation to solve the task quickly.[7]

Second, in the spirit of List (2003, 2004a,b, 2006), we contrast the findings from our natural field experiment with evidence from a second sample of 268 student teams (804 participants) who were paid to perform the same task as part of an economic experiment. These teams were less intrinsically motivated as they did not self-select into the task.[8] We find that despite lower intrinsic motivation, bonus incentives similarly improved performance. Akin to the results from the field experiment, incentives more than doubled the number of teams that managed to solve the task within 45 minutes. As the incentive effect is of similar size, our findings suggest that the efficacy of the bonus incentive does not substantially interact with teams' intrinsic motivation.

Third, our setting furthermore offers us the opportunity to shed light on potential crowding out due to substitution in the spirit of Holmstrom and Milgrom (1991). Teams could request external help when they were stuck by asking for (up to five) hints from ETR staff, which were not relevant for bonus payment eligibility. Interestingly, incentives did not lead to a strong reduction in a team's willingness to explore original solutions among strongly intrinsically motivated customer teams, who self-selected into the task. However, we observe an increase in hint-taking due to incentives among the less intrinsically motivated student teams, who were paid by us to perform the task. Thus, our result highlights an important trade-off regarding substitutional crowding out when teams are not intrinsically motivated to explore on their own.[9]

In a next step, we shed more light on the mechanisms through which incentives operate. To better understand the role of teams' knowledge regarding the production function and potential stake size effects, we exploit natural variation in team size and experience among teams. We find that the efficacy of incentives does not substantially depend on team size but incentives are more effective among experienced teams, suggesting that awareness of how effort translates into performance enhances the positive incentive ef-

[7]Note that surveys among customer teams confirm that their main goal is to achieve success together, and not to stay in the room as long as possible, independent of whether or not a bonus is offered (see also Table A.24).

[8]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.

[9]This interpretation is also in line with findings from additional customer surveys that indicate a strong relationship between own hint-taking behavior and image concerns regarding the latter (see Section 3.3.3 and Figure A.4 in the Appendix).

fect. Further, to study the role of team organization in more detail, we collect additional survey data among student teams. The surveys reveal an increased demand for leadership among treated teams and thus suggest that leadership is an important channel through which performance effects may come about. To uncover the causal role of leadership demand, we then implemented an additional natural field experiment with 281 teams (1,273 participants) in the exact same setting, in which we exogenously varied the demand for leadership. The leadership experiment reveals a substantial positive effect of leadership demand on team performance consistent with the idea that incentives may indeed enhance performance by encouraging team members to seek leadership and take initiative in coordinating and motivating others. As such, we conjecture that the impact of incentives goes beyond merely increasing individual effort. Instead, incentives appear to provide the impetus for teams to endogenously adopt more structured forms of leadership.

Our field experiments encompassed in total more than 5,000 participants and 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 been strongly promoted in the public, for instance by the best selling author Daniel Pink, in a 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 can be a viable instrument to increase performance in such tasks.

To put our findings in perspective, we briefly compare the incentive effects observed in our setting to other field experiments in the literature. In our natural field experiment, the difference in finishing time between treated and control teams amounts to about 0.44 standard deviations. In other work, for routine tasks, performance pay has been shown to improve performance with varying effect sizes (Bandiera et al., 2021). Effects range from zero (Delfgaauw et al., 2020) to 0.90 standard deviations (Hossain and List, 2012).[10] Regarding field experiments involving tasks which are less routine in nature, we can relate our work to research on incentives for teachers and health practitioners. For both

[10]See also Table A.1 in the Appendix and the discussion regarding the retail sector and other settings in Delfgaauw et al. (2020).

professions, typical tasks require rather cognitive than physical effort and may involve (at least sometimes) novel and unknown problems. As such we may consider these settings non-routine and analytical in nature (although it remains unclear if and to what extent complementarities exist). Studies on incentive pay for teachers yield overall mixed results (see, e.g., Fryer et al., 2022), and range from zero effects (Behrman et al., 2015) up to 0.31 standard deviations (List et al., 2018, see also Appendix Table A.1). Evidence regarding incentive pay for health workers is less abundant (Miller and Babiarz, 2014) and observed effects sizes are smaller (see Appendix Section A.1). Regarding other non-routine tasks, we can relate to recent contributions from 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), which also indicate mostly positive incentive effects but almost exclusively measure individual production, instead of joint team production (i.e., in some of these studies, workers may face team incentives but work on individual tasks).[11] One rare exception is a 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 remains unclear whether this null effect is robust, as the authors achieve rather low statistical power.[12] Our work substantially advances this literature by focusing on a jointly solved complex team task and allows for cleanly testing whether and why incentives improve performance. Such settings provide room for incentives to improve team performance not only by increasing workers' effort but they also foster the endogenous emergence of better organizational and leadership structures within teams.

The rest of this paper is organized as follows: Section 2 presents the field setting and the experimental design. Section 3 provides the main results with respect to performance improvements and potential crowding out. We discuss potential mechanisms that shape the efficacy of incentives in Section 4 and provide a more general discussion of our findings in Section 5. Section 6 concludes.

[11] 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.

[12] Ramm et al. (2013) also study individual performance in the candle problem and find no negative incentive effects, whereas Kleine (2021) shows that piece-rate incentives increase the time needed to solve that task.

2 Experimental design

2.1 The field setting

We cooperate with the company *ExitTheRoom*[13] (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 time (one hour) expires, they win—if time runs out before the team solves all quests, the team loses.

A typical escape room usually features several items, such as desks, shelves, telephones, books, and so on. These items may include 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. To illustrate a typical quest in a real-life escape game, we provide a fictitious example.[14] Suppose the participants have found and opened a locked box that contains a megaphone. Apart from being used as a speaker, the megaphone can also play three distinct types of alarm sounds. Among the many other items in the room, there is a volume unit (VU) meter in one corner of the room. To open a padlock on a box containing additional information, the participants will need a three digit code. The solution to this quest is to 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

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

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

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.[15] 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 the 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 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

[15] *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 Appendix Tables A.2 and A.21).

possible is clearly desirable.[16] 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 even 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 either.[17]

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 render escape rooms 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 (268 teams). In particular, it affords a unique opportunity to compare incentive effects among strongly intrinsically motivated teams who have self-selected into the task (regular customers) and incentive effects for less intrinsically motivated teams who were confronted with the task by us and faced the team challenge as part of their paid participation in an economic experiment. The setting's considerable flexibility also enables us

[16]This is also corroborated by additional results from surveys among customer teams confirming that the main goal of teams is to achieve success together (see Appendix Table A.24).

[17]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.

to delve into potential mechanisms by conducting an additional natural field experiment that demonstrates the importance of leadership (see Section 4).

2.2 Experimental treatments and measures of performance

We conducted the field experiment with 3,308 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.[18] In the *Control* condition (238 teams), teams were not offered any bonus.

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[19], 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.[20] 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

[18]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 in Section 2.3.2), 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, they announced the possibility for the team to earn a bonus and had the teams sign a form (see Appendix A.3), indicating that they understood the conditions for receiving 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. To study the role of potential loss aversion akin to Hossain and List (2012), 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. We do not identify major differences across these two conditions and thus pool these treatments in the main text. Additional analyses for these two subtreatments are provided in Appendix Section A.5.4.

[19]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.

[20]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.

treatments with the control condition allows us to estimate the causal effect of bonus incentives on these objective performance measures.

As customer teams are strongly intrinsically motivated to succeed in the team challenge, there is room for potential motivational crowding out. Following Bénabou and Tirole (2003)'s theoretical framework, workers in the incentive condition may interpret incentives as a negative signal (e.g., they may believe that the task is less enjoyable than expected or that incentives are likely implemented for less intrinsically motivated teams). As such, incentives may decrease performance among intrinsically motivated teams. However, our *Bonus45* condition includes an ambitious performance threshold (solving the task in 45 minutes rather than in 60 minutes) which may serve as a positive signal for intrinsically motivated workers. To identify whether the latter may countervail 'negative news' due to incentives or independently affect team performance, we implemented two additional experimental treatments. In the first of these treatments, *Bonus60* (88 teams), we provided the same monetary bonus but did not include the ambitious performance threshold. Instead, the bonus referred to the reference point of 60 minutes (akin to the *Control* condition). That is, teams received the bonus if they completed the task within 60 minutes.[21] In the second 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 but did not pay any bonus.[22] The performance in *Bonus60* as compared to *Control* allows an additional, even stronger test regarding potential motivational crowding in the spirit of Bénabou and Tirole (2003). Differences in performance between *Reference Point* and *Control* further reveal whether referring to an ambitious reference point increases the performance of the teams even if a monetary bonus is absent.[23]

To further evaluate the role of intrinsic motivation for the efficacy of bonus incentives, we exploit the unique opportunity to replicate our (*Bonus45* and *Control*) conditions in a framed field experiment in the exact same setting with different teams that are less intrinsically motivated. For this purpose, we randomly allocated 804 student participants from the subject pool of the social sciences laboratory at the University of Munich (ME-

[21]Akin to the main treatment, we implemented *Bonus60* in two subtreatments, *Gain60* (42 teams) and *Loss60* (46 teams). As treatment differences are again minor, we pool the data in our analysis.

[22]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."

[23]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.

LESSA) into 268 teams. These participants therefore did not self-select into the challenge and were paid to perform the task as part of an economic experiment.[24]

Finally, our setting allows us also to shed light on potential crowding out due to substitution of effort (Holmstrom and Milgrom, 1991), as we are able to analyze incentive effects on teams' requests for external help (i.e., teams use of hints) and to shed light on important channels through which incentives operate.

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.[25] We implemented the main treatments of the field experiment (*Bonus45* 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 *Bonus60* 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.[26]

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. In the natural field experiments, 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 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.

[24]This experiment enabled us to also collect additional data on teams' task perception and team organization (discussed in Sections 3 and 4).

[25]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.

[26]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.

2.3.2 Framed field experiment (student sample)

For the framed field experiment, we invited student participants from the social sciences laboratory at the University of Munich (MELESSA). Between March and June 2016, and January and May 2017, a total of 804 participants (268 groups) took part in the experiment. To avoid selection into the sample based on interest in the task, we recruited these participants using a neutrally framed invitation text that did not explicitly state what activity participants could expect. The invitation email informed potential participants that the experiment consisted of two parts, of which only the first part would be conducted on the premises of MELESSA whereas the second part would take place outside of the laboratory (without mentioning the escape game). They were further informed that their earnings from the first part would depend on the decisions they made and that the second part would include an activity with a participation fee that would be covered by the experimenters.[27]

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.[28] 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

[27] Section A.4 in the Appendix provides a translation of the text of the invitation.

[28] 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.

other at the facilities of ETR. During the performance of the task, the same information about the team performance as in the field experiment was collected. On completion of the task, the participants answered questions about the team's behavior, organization, and their perception of the task individually, on separate tablet computers. At the end, we paid the earnings individually in cash. In addition to the participation fee for ETR, which we covered (given the regular price, this corresponds to roughly €25 per person), participants earned on average €7.53, with payments ranging from €3.50 to €87.[29]

3 Results

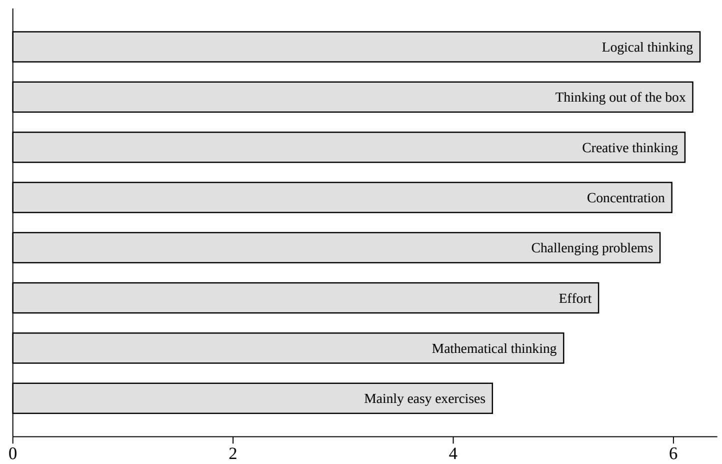
3.1 Task perception and randomization

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 establish internal validity of our experimental appraoch without interfering with the standard procedures at ETR, we could not run extensive surveys and, e.g., ask customer participants of our natural field experiments about their perception of the task. However, we asked the student participants from the framed field experiment ($n = 804$) to what extent they agree that the team task exhibits various characteristics (using a seven point Likert scale).

Figure 1 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,

[29]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.



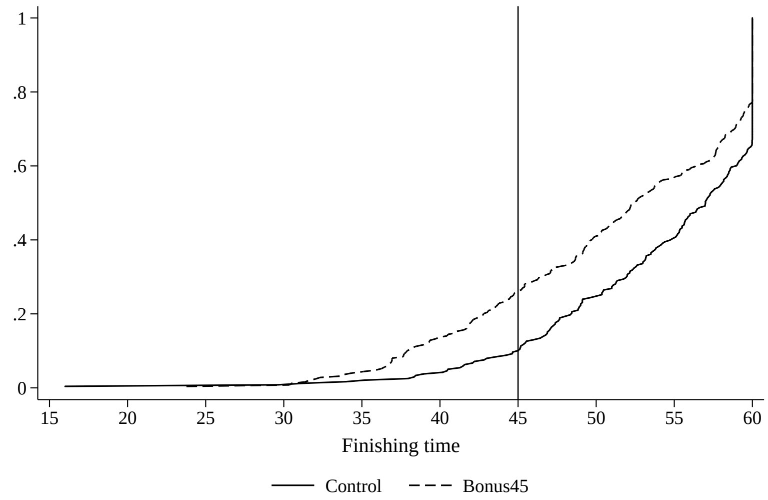
Notes: The figure shows mean answers of $N = 804$ student participants to eight questions concerning attributes of the task. Answers were given on a 7-point Likert scale.

Figure 1: Task perception

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.



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 2: Finishing times in *Bonus45* and *Control* in the field experiment

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).[30] We therefore provide results from both the regression specifications without controls and the regression specifications in which we control for the estimated age ranges (and other observables).

3.2 Bonus incentives and team performance

We now turn to our primary research question, whether providing bonus incentives improves 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 for task completion. Figure 2 shows the cumulative distribution of finishing times with and without bonus incentives in the field experiment. The vertical line marks the time limit for receiving 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 noisy 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

[30] This does not change when adjusting for multiple hypothesis testing according to List et al. (2019).

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

	Probit (ME): Completed in less than 45 minutes			
	(1)	(2)	(3)	(4)
Bonus45	0.165***	0.164***	0.188***	0.151***
	(0.024)	(0.022)	(0.025)	(0.041)
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	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.5 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.

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 $= 0.01$). Adjusting p-values for multiple hypotheses testing as suggested in List et al. (2019) yields similar results. For further details, see also Table A.8 and Section A.10.1 in the Appendix.

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 treatment *Bonus45*. Bonus incentives are estimated to increase the probability of completing the task in less than 45

Table 3: Influence of main bonus treatment on hazard rates

	Cox Proportional Hazard Model: Finishing the Task					
	First 45 min			Last 15 min		
	(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.

minutes by 16.5 percentage points. This effect is substantial and equivalent to expanding team size from 4 to 6 members. In Column (2), we add observable team characteristics.[31] 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. In Appendix Table A.6, 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.

Since the incentive only rewards completing the task in the first 45 minutes, it should become ineffective for 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 investigate these conjectures in more detail, we run a Cox proportional hazard model, where we define the hazard as completing the task. If our prior was true, we should observe the treatment to have a 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.

[31]From the set of characteristics in these and the following analyses, group size, experience with escape games and the share of males in a team have a positive effect on performance, whereas English-speaking groups perform slightly worse. For more details see Table A.3, Column (1).

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 does not 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 have caused strong feelings of disappointment resulting in substantially worse performance after teams failed to achieve the threshold relevant for the bonus payment in our setting. We conclude:

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

3.3 Potential crowding out of intrinsic motivation

Importantly, the results from our field experiment demonstrate that bonus incentives substantially improve team performance among strongly intrinsically motivated teams. As such, the monetary reward of the bonus appears to outweigh potential negative effects due to the crowding out of intrinsic motivation. However, the bonus incentive in *Bonus45* was tied to an ambitious performance threshold (45 minutes) that only ten percent of teams in *Control* were able to achieve. In turn, arguing in the framework of Bénabou and Tirole (2003), teams in our treatment condition may have also interpreted the bonus as a positive signal about their own ability as the ambitious threshold signals they can achieve an extraordinary performance. Hence, we consider it crucial to investigate whether bonuses also work if they are not coupled with ambitious performance thresholds. Furthermore, we desire to explore the robustness of incentive effects among a sample of less intrinsically motivated teams. Doing so allows us to go beyond an analysis of the potential 'net effect' of incentives and potential crowding out. In particular, if we observe similar effect sizes among differently intrinsically motivated teams, the net effect likely coincides with the 'pure' positive effect of bonus incentives. Finally, we seek to uncover whether crowding out can be observed in the form of substitution of (multi-dimensional) effort by shedding light on teams' exploration behavior (i.e., hint-taking).

3.3.1 Ambitious performance thresholds and incentives

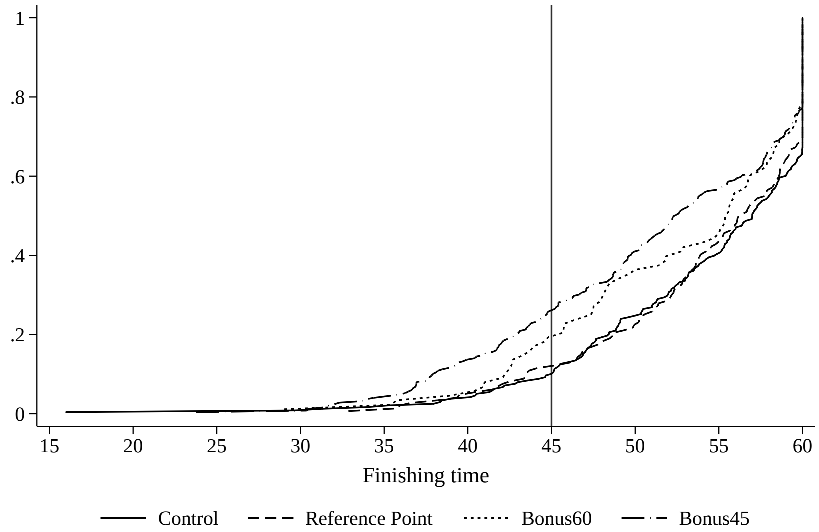
To understand whether ambitious performance thresholds countervailed a potential crowding out of intrinsic motivation or independently caused positive performance effects, we conducted two additional treatments. In *Bonus60*, we paid a bonus for completing the task in 60 minutes.[32] In *Reference Point*, we introduced the 45-minutes threshold as a salient reference point but did not pay a reward. Figure 3 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, also when they are coupled with a less ambitious performance goal of 60 minutes (*Bonus60* vs. *Control*, Mann–Whitney test, p-value $= 0.05$; *Bonus45* vs. *Control*, Mann–Whitney test, p-value < 0.01, with *Bonus45* vs. *Bonus60*, Mann–Whitney test, p-value $= 0.24$). Further, we do not observe that the ambitious reference point independently improves as the cumulative distribution of remaining times in *Reference Point* almost perfectly overlaps with the cumulative distribution function in *Control* (Mann–Whitney test, p-value $= 0.78$).[33]

For completeness, we provide a regression analysis for the full sample of ETR customer teams in Table 4. 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.[34] It also becomes apparent that this finding is robust to the addition of covariates and fixed effects. Moreover,

[32] As with *Bonus45*, we again split the treatment into two sub-treatments (framing the bonus as a gain or a loss). Similar to the analysis for *Bonus45*, we do not differentiate between the gain and the loss frame of *Bonus60* in the following. No difference between the frames in *Bonus60* emerged.

[33] 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.10.1.

[34] Appendix Table A.7 confirms these findings for remaining time as the dependent variable.



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 3: Finishing times for all treatments in the field experiment

Table 4: 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 *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

a post-estimation Wald test rejects the equality of coefficients of *Bonus60* and *Reference Point* in all specifications (Columns (1) to (4), p-values < 0.1). Similarly, the coefficient of *Bonus45* is significantly larger than the coefficient of *Reference Point* in all specifications (p-value $= 0.07$ in Column (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 (Column (2), p-value $= 0.095$). We summarize this finding in Result 2:

Result 2 *Bonuses with less ambitious performance thresholds do not lead to additional motivational crowding out. Introducing an ambitious reference point (indicating extraordinary performance) alone is not sufficient to induce a performance shift.*

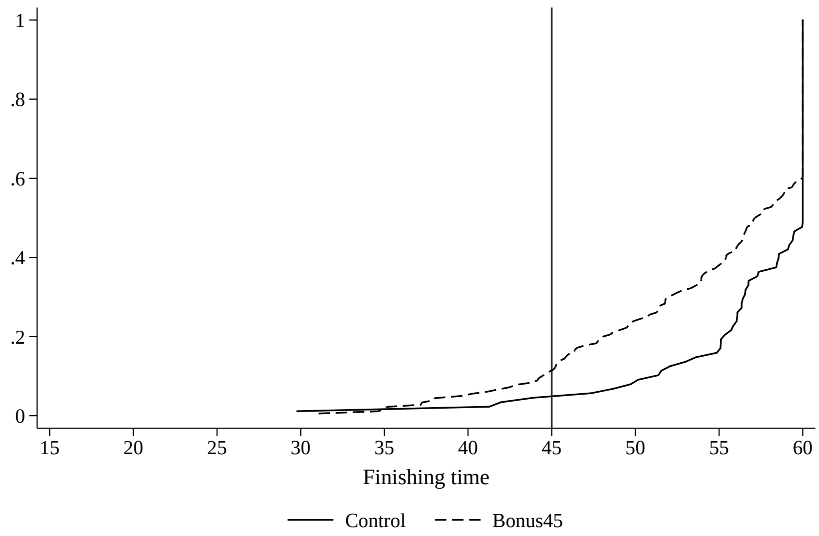
3.3.2 Incentive effects among less intrinsically motivated teams: Results from the framed field experiment

To test whether the performance-enhancing effect of bonus incentives is also present in teams other than the self-selected 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 ETR customers self-select into the task and are likely to be intrinsically motivated to perform well, student teams from the laboratory subject pool are assigned 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.[35]

We randomized 268 teams of three students into treatments *Control* (88) and *Bonus45* (180).[36] Across both treatments, teams do not differ significantly in any observed characteristics. The average share of males in *Bonus45* (0.43) is not significantly different to *Control* (0.45) (Mann-Whitney test, p-value $= 0.31$), and neither is the share of teams with at least one experienced member (0.36 versus 0.36, χ^2 test, p-value $= 0.17$), or teams'

[35]Teams of ETR customers were also formed endogenously and varied in size, whereas we randomly assigned students to teams of three participants. Further, student teams differ along observable dimensions, such as age, gender, and experience with the task. They were on average younger (23.03 years), 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 escape game experience).

[36]Akin to our analyses regarding the natural field experiment, we pool the two subtreatments *Gain45* (90) and *Loss45* (90) also for the student teams, and provide additional results on the framing of incentives in Appendix Section A.6.



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

Figure 4: Finishing times in *Bonus45* and *Control* in the framed field experiment (student sample)

median age (22.96 versus 23.18, Mann-Whitney test, p-value $= 0.72$). 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 4 shows the performance of teams in the framed field experiment and is the student sample analogue to Figure 2. 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 large 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 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).[37] These results are also robust to adjusting p-values for multiple hypotheses testing as suggested in List et al. (2019) (see Appendix Section A.10.2 for more details).

In addition to the non-parametric tests, we run regressions analogously to the analysis for the customer sample. As before, we control for the share of males in a team, average age and experience with escape games.[38] Table 5 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. Also for the student sample, the positive effect of bonus incentives is reflected qualitatively in the analyses of the time remaining (see Appendix Table A.10). These results, coupled with our findings from the natural field experiment, emphasize that a crowding out of intrinsic motivation does not seem to distort the 'pure' effect of incentives. We conclude:

Result 3 *Teams' intrinsic motivation does not substantially alter the efficacy of incentives.*

3.3.3 Bonus incentives and teams' willingness to explore

In addition to studying crowding out within our performance outcome, our setting also allows us to shed light on crowding out in other effort dimensions (Holmstrom and Milgrom, 1991). In particular, bonus incentives may alter teams' intrinsic motivation to explore original solutions. In fact, previous research has suggested that performance-based financial incentives may affect workers' willingness to explore in an experimentation task (see, e.g., Ederer and Manso, 2013). By design, if teams request external help by asking for hints, they explore less on their own. In the following, we analyze how many out

[37] Appendix Table A.9 summarizes these findings and provides further details with respect to the framing of incentives.

[38] 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 5: Probit regressions: Completed in less than 45 minutes (student sample)

	Probit (ME): Completed in less than 45 minutes			
	(1)	(2)	(3)	(4)
Bonus45	0.075*	0.073*	0.075*	0.079**
	(0.042)	(0.041)	(0.039)	(0.036)
Fraction of control teams completing the task in less than 45 min	0.05	0.05	0.05	0.05
Control Variables	No	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes
Week Fixed Effects	No	No	No	Yes
Observations	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.11 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.

of the five possible hints teams request under the different treatment conditions, as well as whether they are more likely to take hints earlier in the presence of incentives.[39]

Table 6 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) significant increases in hint taking in the bonus treatments as compared to *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 5. 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 5 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

[39]In Appendix Section A.9, we provide additional evidence that the increase in hint-taking in the framed field experiment is unlikely due to increased importance of risk aversion when incentives are in place.

Table 6: Hints requested in the field experiment and the framed field experiment

	Control	*Bonus45*
within 60 minutes		
Field Experiment (487 groups)	2.92 (1.55)	3.10 (1.34)
Framed Field Experiment (268 groups)	3.74 (1.04)	4.11 (0.98)***
within 45 minutes		
Field Experiment (487 groups)	1.97 (1.22)	2.36 (1.15)***
Framed Field Experiment (268 groups)	2.33(0.93)	3.17 (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.

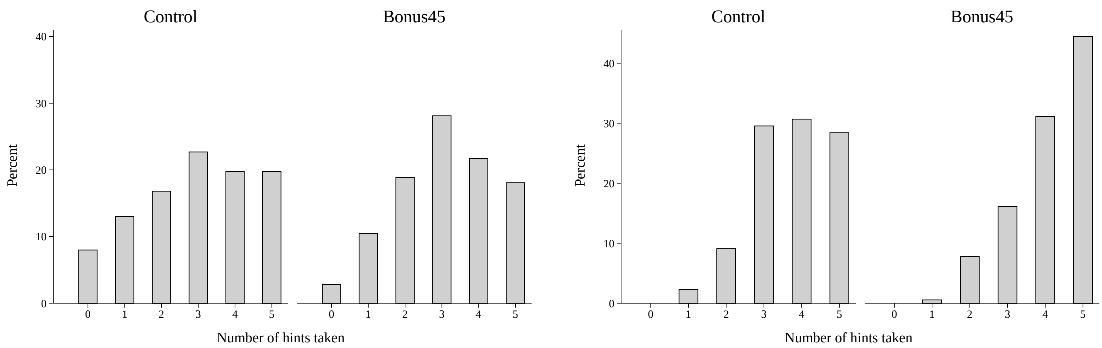
take hints when facing incentives (χ^2-test, p-value=0.029). Roughly 75 percent of these teams take four or five hints when facing incentives, as compared to 59 percent doing so in *Control*. Regression analyses for hint taking including additional controls (see Table 7, Columns (1), (2), (5), and (6)) confirm these results.[40]

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 7). When we add additional controls and fixed effects (Columns (4) and (8) of Table 7), 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 intricate way when teams can choose whether or not to explore original and innovative solutions on their own.[41] Incentives strongly reduce the willingness to explore original solutions of teams that did not self-select into the task, highlighting a substitution of effort due to incentives in line with the multitasking framework by Holmstrom and Milgrom (1991). However, such substitution is much less

[40] An ordered probit regression yields qualitatively similar results, see Appendix Table A.13.

[41] These findings complement recent work on incentive effects in meaningful routine tasks (Kosfeld et al., 2017).



(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 5: Hints requested across samples and treatments

Table 7: OLS regressions: Number of hints requested

	OLS: Number of hints requested							
	Field Experiment				Framed Field Experiment			
	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) are 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.

prevalent among strongly intrinsically motivated customer teams indicating that these teams may derive utility from making progress on their own and hence take fewer hints. To understand whether this idea is reflected in teams' perceptions, we ran an additional survey among ETR customers ($n = 201$) and analyze how teams' perceptions differ conditional on their own hint taking behavior. While, both, teams that take few (less than three) or many hints (three or more) similarly agree that hints are used to find solutions to difficult puzzles (χ^2 test, p-value$= 0.71$), we observe that teams taking few hints perceive hint taking more negatively, in particular as less creative (χ^2 test, p-value < 0.01), less original (χ^2 test, p-value < 0.01) and less fun (χ^2 test, p-value < 0.01).[42] An alternative explanation for reduced substitution among intrinsically motivated teams (as compared to hired teams) can be found in the framework of Bénabou and Tirole (2003). Here, strongly intrinsically motivated teams may wish to compensate potential 'negative news' about their ability due to incentives and thus not substitute exploration effort for hints when incentives are in place. However, the latter should likely result in less hint-taking among teams in the bonus condition as compared to *Control* (which we do not observe). Further, among the intrinsically motivated customer teams we see no significant differences in the number of hints taken when bonuses are combined with more ambitious (as compared to less ambitious) performance thresholds (3.09 hints in *Bonus45* versus 3.26 hints in *Bonus60*, Mann Whitney, p-value=0.3271) rendering compensating behavior unlikely. 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, but have not yet established *how* incentives improve performance. We attempt to provide insights on likely mechanisms by pursuing two distinct avenues. First, to gain a better understanding of what distinguishes teams that do respond to incentives from those who do not, we discuss whether any particular observable team features

[42]For further details on the survey, see Appendix Section A.14.

interact with the observed efficacy of incentives. Second, we investigate how incentives affect behavior, particularly team organization. Post-experimental survey responses of more than 260 student teams identify increased demand of leadership as a potential channel, which we subsequently investigate in an additional natural field experiment.[43]

4.1 When do incentives work?

We first investigate whether the efficacy of incentives interacts with observable team characteristics within 45 minutes (Appendix Table A.3) as well as teams' remaining times (Appendix Table A.4). Regarding teams' propensity to finish within 45 minutes, we do not find significant interactions with the teams' gender share, team size, teams' language, or whether teams participated as part of a company event. This seems to indicate that bonus incentives appear to be similarly effective for teams of different size and levels of diversity. We further investigate whether teams with experienced team members react differently to incentives than inexperienced teams. Experienced members possess more knowledge about how team effort transforms into team success, which could enhance the effects of incentives. We find a positive, economically and statistically significant interaction of bonus incentives and experience. We estimate that the positive bonus effect is about 1.5 times larger for experienced teams (see Column (4) in Appendix Table A.3). This suggests that a good understanding of the production function is crucial in this setting for harnessing the benefits from incentives. The latter is also mirrored in teams' remaining times where the bonus tends to be more effective for experienced teams, even though not at conventional significance levels (p-value = 0.10, see Column (4) in Appendix Table A.4). For remaining times, we also find that a higher share of men relates positively to performance but decreases the effectiveness of incentives (possibly due to ceiling effects). Similarly, when studying the efficacy of incentives across predicted performance quintiles (based on observable team characteristics), we find weaker incentive effects for teams predicted to perform very well (see Appendix Figure A.1). This result is in line with the notion that the efficacy of incentives can be weaker for teams that already exert high levels of effort. Notably, we do find robust, positive, and significant incentive effects among all other quintiles. Finally, and akin to the analyses regarding the proba-

[43] Additionally, in Appendix Section A.12, we provide a broader discussion of the dimensions along which incentives may change behavior within teams including additional surveys and an additional laboratory experiment.

bility of finishing within 45 minutes, we find that the efficacy of incentives for improving remaining times does not significantly differ for the number of team members, whether the team is English or German speaking, or whether the team challenge was booked by a company or private team. We summarize these findings in Result 5:

Result 5 *Understanding of the production function enhances incentive effects.*

4.2 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 8 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 8 shows that incentives overall do not strongly affect agreement with the statements we provided. However, teams appear to be notably more stressed when facing incentives than teams in *Control* (Mann–Whitney test, p-value < 0.01).[44] 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-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 8 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

[44] We are agnostic about whether this increase in stress levels is a direct result of incentives or a by-product of increased effort levels.

Table 8: Answers to post-experiment questionnaires

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

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. The latter result is also reflected in an alternative approach to adjust for multiple hypothesis testing by using principal component factor analyses for dimensionality reduction following the Kaiser-Guttman rule (see Loehlin and Beaujean, 2016). We do so separately for Questionnaire 1 and 2 in Appendix Table A.12. For Questionnaire 1, the analysis retains five factors. We name these factors 'general team collaboration' (Factor 1), 'team cohesion' (Factor 2), 'dominance' (Factor 3), 'documentation' (Factor 4) and 'intensity' (Factor 5).[45] We find that general team collaboration does not significantly differ across treatments (Mann-Whitney test: p-value=0.76), and neither does dominance (Mann-Whitney test: p-value=0.11), whereas incentives tend to increase team cohesion (Mann-Whitney test: p-value= 0.07), and intensity (Mann-Whitney test: p-value$<$0.01) but decrease documentation (Mann-Whitney test: p-value=0.02). Regarding Questionnaire 2, we retain three factors which we term: 'cooperative' (Factor 1), 'leadership' (Factor 2), and 'struggling' (Factor 3).[46] Cooperative behavior (Factor 1) does not significantly differ across treatment conditions (Mann-Whitney test: p-value= 0.34). Leadership (Factor 2) is significantly more pronounced with incentives (Mann-Whitney test: p-value$<$0.01). 'Struggling' in teams (Factor 3) tends to be lower with incentives, but statistically insignificantly so (Mann-Whitney test: p-value=0.26). Overall, both analysis indicate that incentives appear to change team organization and stimulate the demand for and the emergence of leadership.

4.3 The causal effect of leadership

Having established that bonus incentives strengthen the demand for leadership and may have led to the endogenous emergence of a team leader, we consider it crucial to provide evidence for a causal link between (an increased demand for) leadership and performance.

[45]Items from Questionnaire 1 that load heavy on Factor 1 are : (5) (6), (7), (9), (13), (15), and (18). The items loading heavy on Factor 2 are: (8), (10), (12), (16), (17), and (19). The items loading heavy on Factor 3 are: (2) and (4). The items loading heavy on Factor 4 are: (3), (11). The items loading heavy on Factor 5 are: (1), (14).

[46]Items that load high on Factor 1 are: (1, negatively), (5), (7), (8), (10), (11), and (12). Items that load high on Factor 2 are: (2) and (6). Items that load high on Factor 3 are: (3), (4), and (9).

To do so, we conducted another natural field experiment in the same setting. Between January and March 2018, 1,273 regular customers in 281 teams were assigned to one of two experimental conditions: *Control-L*, and *Leadership*. As in our *Control* conditions reported earlier, participants in *Control-L* did not experience any intervention. To participants in *Leadership*, however, ETR staff highlighted the importance of leadership to succeed in the task, and encouraged them to select a leader according to a short standardized script:

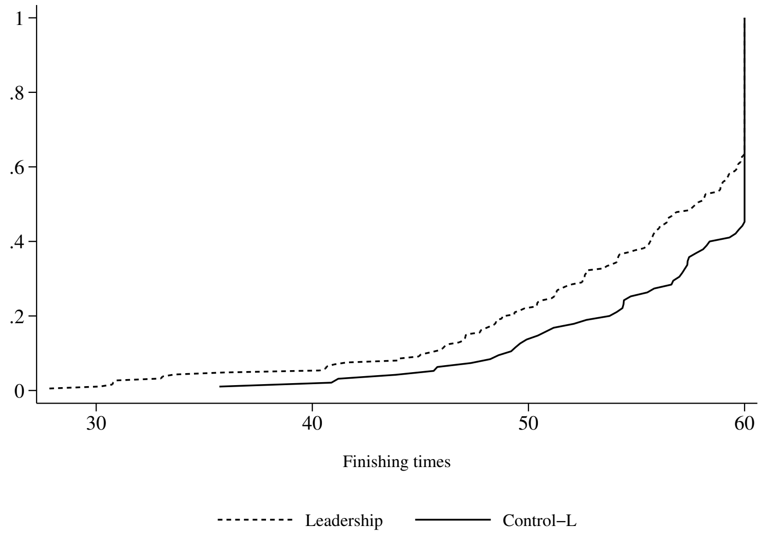
> "One piece of advice before you begin: a good team needs a good leader. Past experience has shown that less successful teams often wanted to have been better led. Thus, decide on someone of you, who takes over the leading role and consistently motivates/coordinates the team."[47]

Figure 6 shows the cumulative distribution functions (CDFs) of finishing times across conditions. Teams in the *Leadership* treatment condition clearly perform better than those in the *Control-L* condition. Specifically, 63% of teams finish the task within the time limit of 60 minutes in Leadership, whereas only around 44% do so in Control (Pearson χ^2 test: $p < 0.01$). In addition to being more likely to complete the task, teams that were encouraged to choose a leader also solve the task faster (average remaining times: 3m10s in *Control-L*, 5m29s in *Leadership*, Mann-Whitney test: $p < 0.01$).

These non-parametric results are confirmed by a series of Probit regressions, in which we step-wise introduce additional control variables and fixed effects as in Table 2. In Table 9, we estimate the average marginal effect of *Leadership* on the probability to complete the task within 60 minutes. As before, we cluster standard errors at the daily level, which is also the level of random treatment assignment. In all specifications, we find that exogenously shifting the demand for *Leadership* significantly increases teams' probability to succeed within 60 minutes. The estimated average marginal effect amounts to an increase of 17 percentage points as compared to *Control-L*, implying a relative increase in the fraction of successful teams of about 38%. In Appendix Table A.20, we present the analyses for the remaining time. The implied average marginal effects show that raising awareness of the importance of leadership demand unambiguously increase the remaining time

[47] The treatment *Leadership* consisted of two subtreatments that differed only by whether the last sentence stressed the word "motivate" or "coordinate". As the effects of stressing different leadership functions are not the focus of the present paper, details are relegated to Englmaier et al. (2021).

Figure 6: Leadership: CDFs of finishing time



Notes: The figure shows the cumulative distribution of finishing times for teams in (*Leadership*) and (*Control-L*).

upon task completion by on average 2m48s.[48] These findings, coupled with the survey evidence that incentives increased the demand for leadership, show that the resulting emergence of leadership mediates the positive effects of incentives on performance. We summarize our findings in Result 6:

Result 6 *Bonus incentives induce demand for leadership and alter team organization. Exogenously shifting the demand for leadership results in substantial performance improvements.*

[48]Note that the magnitudes are hardly comparable with the results presented in Tables 2 and A.6, as incentives targeted task completion after 45 minutes, whereas the leadership intervention only targeted completion at the 60 minute mark. The cleanest comparison for the case of incentives would be to regress remaining times or likelihood of completion in 60 minutes on the *Bonus60* treatment. Doing so in the full specification results in a marginal effect of an additional 2m44s remaining time and 12.5 percentage points higher completion probability. The latter is somewhat lower, albeit not significantly so, than the effect of leadership, but note that teams in the different control groups were differently successful (0.442 in *Control-L* versus 0.67 in *Control*). Thus, leadership possibly had a larger scope to improve performance on the extensive margin. Therefore, the emergence of leadership seems to exhibit a comparable potential to improve performance as offering bonus incentives.

Table 9: Probit regressions: Leadership: Completed in less than 60 minutes

	Completed within 60 Minutes			
	(1)	(2)	(3)	(4)
Leadership	0.182***	0.187***	0.185***	0.168***
	(0.051)	(0.052)	(0.065)	(0.051)
Fraction of teams in *Control-L* completing the task in 60 minutes	0.442	0.442	0.442	0.442
Controls	No	Yes	Yes	Yes
Staff FE	No	No	Yes	Yes
Week FE	No	No	No	Yes
Observations	281	281	281	281

Notes: The table displays average marginal effects from Probit regressions of whether a team completed the task within 60 minutes on our *Leadership* indicator (with *Control-L* as base category). Each column indicates whether team controls (group size, share of male participants, experience with escape games, median age, language spoken, private versus team-building events, actively taken walkie-talkie), staff, and week fixed effects are included. Standard errors in parentheses are clustered at the daily level, with significance levels * = $p < 0.10$, ** = $p < 0.05$, and *** = $p < 0.01$.

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 10, 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$= 0.50$). Priors of $\pi > 0.10$ 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.

Table 10: Post-Study Probabilities

Achieved power for...	χ^2-tests on success dummy (45 & 60 mins) in framed field	χ^2-tests on success dummy (45 mins) in nat. field	χ^2-tests on success dummy (60 mins, nat. field) and t-tests on remaining time (field and framed field)
	(1)	(2)	(3)
	0.45	0.70	0.95
Prior probability	Posterior	Posterior	Posterior
0.05	0.32	0.42	0.5
0.1	0.50	0.61	0.68
0.2	0.69	0.78	0.83
0.4	0.86	0.90	0.93
0.6	0.93	0.95	0.97
0.8	0.97	0.98	0.99
0.9	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.

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.

To quantify a reasonable prior, in March 2023, we asked 400 participants from a pool of HR experts by survey provider Cint for their priors on the effectiveness of incentives in non-routine analytical team tasks. Half of the HR experts ($n = 203$) were asked about the effectiveness of bonus incentives in escape challenges.[49] We explicitly informed these experts about the nature of the task at hand and asked them to guess how many out of one hundred teams i) would become faster, ii) would become slower, and iii) would neither, once they received the opportunity to earn a bonus. On average, HR experts believed that 40.38% of teams would improve in performance, 23.33% of teams would decline, and

[49]For the other half, we elicited expectations regarding abstract non-routine tasks. We present differences between these expectations below, when discussing external validity drawing on the SANS conditions provided in List (2020).

outcomes for 36.29% of teams would remain unchanged. As Table 10 shows, a prior of approximately 0.4 (believing a positive effect is less likely than a coin flip) in all cases enables posteriors close to believing a true effect to exist.[50] These calculations emphasize the strong updating that decision makers should undergo as they learn about the results from our study.[51]

Our series of large-scale field experiments constitutes, to the best of our knowledge, the first systematic investigation into incentive effects in non-routine analytical and jointly solved team tasks. To discuss external validity of our results, we consider it useful to draw on the SANS conditions introduced in List (2020): *Selection, Attrition, Naturalness,* and *Scaling*.[52] Our two samples reported in this paper consist of actual customers of ETR, as well as students, who conceivably differ along many dimensions.[53] As our documented treatment effects carry over to participants from both samples, this seems to indicate that *selection* is not a primary concern. Additionally, university students are likely (on average) similar to workers in many non-routine, analytical team work environments, as these frequently require higher levels of education. Neither do we deem *attrition* to be a major concern, as none of the participants opted out from our framed field experiment nor were participants aware of being studied in the natural field experiment (and hence selective attrition could not occur in the latter either). In terms of *scaling*, we wish to note that stakes in our setting are substantially lower than typical bonuses paid in firms. On the other hand, our results in Appendix Tables A.3 and A.4 have not shown a significant interaction between incentives and team size, suggesting that, at least locally, the size of the incentive is less important. As such, we would expect

[50] As HR experts in the survey could have believed that improving teams became substantially faster, whereas declining teams only moderately slower, we also asked for the number of minutes teams would be expected to be faster/slower (conditional on being faster/slower). The small difference (48 seconds) between the two is not statistically significant, Wilcoxon signed-rank test $p = 0.2518$.

[51] In addition to HR experts, in Appendix Section A.12, we describe a survey with two samples: i) a hand-curated list of academics working in personnel economics, and ii) respondents from "ESA discuss", a mailing list for academic experimental economists. We asked both samples if they believed incentives to influence performance in non-routine analytical team tasks. Over 80% reported to think that incentives have at least some positive effect. A 0.4 prior for HR experts therefore seems to be a lower bound among relevant samples, pushing the posterior potentially even closer to certainty.

[52] For similar applications of this approach see also Holz et al. (2023), Goldszmidt et al. (2020), and Fehr et al. (2022).

[53] As we do not collect background information about customers apart from age, we can only assume that not all ETR participants are University-educated (and different along the many margins that typically correlate with this). In light of comparatively low rates of University attendance in Germany of below 30%, we deem this assumption reasonable. Any differences in characteristics may be in addition or give rise to differences in preferences, constraints, and beliefs (e.g., differing levels of intrinsic motivation for the task).

to observe, if anything, larger effects when applying our interventions in various work environments.[54]

In terms of *naturalness*, we concede that our task indeed is only *one* example of a non-routine analytical team task. Given the vast number of work environments that fall under this broad classification, other jobs may contain additional idiosyncratic features that could influence the presence of the effect we detect. But importantly, our task, and all other non-routine, analytical team tasks share three features: 1) they are non-routine, 2) require analytical thinking, and 3) are conducted in teams. Building our experiment on these commonalities ensures that our analysis covers the essence of this class of tasks. This is corroborated by a survey among 400 HR experts we conducted in February 2023.[55] 203 HR experts were asked to report how many out of 100 teams they expected to perform better, worse, and the same when offered a monetary bonus for completing an escape room challenge. Another 197 HR experts reported the same numbers for abstract non-routine analytical team tasks (without mentioning escape games). Across both settings, HR experts believe incentives to be similarly effective. They predict that 41.37% of teams will improve for abstract non-routine tasks versus 40.38% for escape challenges (p-value $= 0.6638$, Mann-Whitney test). 21.48% versus 23.33% of teams are predicted to perform worse (p-value $= 0.4107$, Mann-Whitney test), and 37.15% versus 36.29% similarly (p-value $= 0.7990$, Mann-Whitney test). While we argue, based on these insights, that additional idiosyncratic features of other tasks should not constitute a major threat to external validity per se, we wish to discuss idiosyncratic features of our task one by one.

First, ETR customers choose to perform the team challenge and are willing to incur costs in order to do so. This suggests that they are likely to receive some utility from performing the task (e.g., they are motivated by the challenge of solving puzzles and tackling different angles of the complex task), which may not hold in general 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 challenging non-routine tasks in their job.[56] Naturally, self-selection into

[54]As we observe consistent effects of incentives across both samples (which may have very different costs and benefits), the use of incentives seems to be scalable to a large number of cases that vary along similar dimensions.

[55]The survey is described in detail in Appendix Section A.13.

[56]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

work environments with non-routine tasks will become less important as current labor market trends continue, as likely many jobs will transform and include more non-routine team elements in the future. Importantly, as we find very similar effects of incentives on teams' finishing times across both of our 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. 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 deliver. As the escape game does not feature such greater goals, 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 assigned 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. Further, our results highlight that the positive incentive effects mainly stem from improved organization and more structured leadership, which should also benefit teams performing tasks with greater goals. Finally, we consider our finding of broad applicability, as many workers perform non-routine tasks in occupations that do not necessarily serve greater goals.

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. We agree that some non-routine tasks may feature open solutions. However, we do not perceive it as a threat to external validity for two reasons. First, many complex problems of interest arguably have only a single (optimal) solution, but there exist multiple ways of arriving at that solution, both in the workplace, 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

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. Further, escape challenges are also used in the context of hiring, where employers can use team based approaches to screen future employees' skills to work in non-routine tasks (https://www.eseibusinessschool.com/experimental-escape-room-recruitment-event-esei-tradler/).

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. Second, while in our setting, the optimal solution is known to the creators of the escape challenge, it is unknown to the teams taking part. Throughout the task, teams may not know if there exists only one solution to each sub-problem, or if picking one out of a number of possible solutions will let them advance in the task.

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 could thus estimate inflated incentive effects. However, as the task requires mainly cognitive effort, observability of co-workers' effort provision is limited in our setting too. 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 strongly intrinsically motivated (customer sample) and less strongly intrinsically motivated teams (student sample). However, incentives increase performance in both samples to a similar degree.

Finally, we would 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 long 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 jointly solved non-routine analytical team tasks in the light of widespread claims of "if-then rewards" being ineffective in such modern tasks (Pink, 2009, 2011, in the nomenclature of List (2020), we thus view the findings as WAVE1 insights). Hence, while we do provide robust evidence in a controlled field setting and from two distinct samples that incentives do improve team performance, more replications will need to be completed to understand if the size of the result applies to other non-routine tasks and occupational environments.

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. We find an economically and statistically significant positive effect of incentives on performance: Teams are more than twice as likely to complete the task in 45 minutes under the incentive condition than under the control condition, and the difference in finishing time between treated and control teams amounts to about 0.44 standard deviations observed in control.

Our comprehensive approach further allowed us to isolate important channels through which incentives may operate in jointly solved non-routine analytical team tasks. First, as these tasks are often performed by intrinsically motivated teams, we studied whether incentives lead to crowding out. Following the framework of Bénabou and Tirole (2003), in which crowding out occurs because incentives are perceived as negative signals about the task or teams' ability, we studied the efficacy of bonuses among teams that were intrinsically motivated to succeed in the task at hand. We varied whether bonuses were coupled with less or more ambitious performance goals and find a substantial improvement in teams' performance in both conditions. Thus, we document a robust 'net' positive effect of bonus incentives that renders crowding out in the spirit of Bénabou and Tirole (2003) unlikely. Further, and in line with the latter interpretation, we find that bonus incentives lead to similar performance improvements among intrinsically motivated (customer) teams who self-selected into the task and less intrinsically motivated (student) teams that were assigned to perform the task. However, our experiments still document an important trade-off related to crowding out in the form of substitution of effort (Holmstrom and Milgrom, 1991). Particularly among teams that we assigned to perform the task, we found a reduction in teams solving problems on their own and taking more hints.[57]

[57]Note that there are 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,

Second, in contrast to routine tasks, in which the relationship between effort is often deterministic, non-routine analytical team tasks are characterized by a noisier relationship between effort and performance. As such, teams' productivity may depend on how individual efforts are combined and teams' understanding of the production function may shape the efficacy of incentives. In line with this idea, we find that incentives are most effective for experienced teams rendering understanding of the production function a crucial mediator for the efficacy of incentives in non-routine tasks.[58] Other team-specific factors that could contribute to the efficacy of incentives (e.g., team size) turn out to be less important. Further, we document that incentives induce important changes in team organization and increase teams' demand for leadership. As such, incentives may not only fulfill their required function to increase performance but also provide benefits beyond this, by fostering more structured leadership within teams, which can causally improve team performance. Finally, we find that teams in the incentive condition reported to be significantly more stressed. Although, in our setting, we did not observe that increased stress levels reduced teams' willingness to perform similar tasks again, in general firms may worry that increased stress may result (in the long run) in costly turnover. Overall, our findings thus emphasize robust positive effects of bonus incentives but highlight also important trade-offs between employee production and turnover as well as regarding potential crowding out in the form of substitution (in our setting exploration vs. hint taking), particularly when teams are less intrinsically motivated to explore on their own.

Taken together, our results raise several 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

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 significantly better performance by teams facing incentives but no significant differences in hint taking (see Appendix Table A.14).

[58]The latter finding also renders the idea 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 unlikely.

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.

An, L. C., Bluhm, J. H., Foldes, S. S., Alesci, N. L., Klatt, C. M., Center, B. A., Nersesian, W. S., Larson, M. E., Ahluwalia, J. S., and Manley, M. W. (2008). A randomized trial of a pay-for-performance program targeting clinician referral to a state tobacco quitline. *Archives of Internal Medicine*, 168(18):1993–1999.

Angrist, J., Lang, D., and Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 1(1):136–163.

Angrist, J. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review*, 99(4):1384–1414.

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.

Bandiera, O., Fischer, G., Prat, A., and Ytsma, E. (2021). Do women respond less to performance pay? building evidence from multiple experiments. *American Economic Review: Insights*, 3(4):435–454.

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.

Bardach, N. S., Wang, J. J., De Leon, S. F., Shih, S. C., Boscardin, W. J., Goldman, L. E., and Dudley, R. A. (2013). Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: a randomized trial. *JAMA*, 310(10):1051–1059.

Barlevy, G. and Neal, D. (2012). Pay for percentile. *American Economic Review*, 102(5):1805–1831.

Barrera-Osorio, F. and Raju, D. (2017). Teacher performance pay: Experimental evidence from Pakistan. *Journal of Public Economics*, 148:75–91.

Basinga, P., Gertler, P. J., Binagwaho, A., Soucat, A. L., Sturdy, J., and Vermeersch, C. M. (2011). Effect on maternal and child health services in rwanda of payment to primary health–care providers for performance: an impact evaluation. *The Lancet*, 377(9775):1421–1428.

Behrman, J. R., Parker, S. W., Todd, P. E., and Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: Results from a social experiment in mexican high schools. *Journal of Political Economy*, 123(2):325–364.

Bénabou, R. and Tirole, J. (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies*, 70(3):489–520.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

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

Boly, A. et al. (2011). On the incentive effects of monitoring: evidence from the lab and the field. *Experimental Economics*, 14(2):241.

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

Brownback, A. and Sadoff, S. (2020). Improving college instruction through incentives. *Journal of Political Economy*, 128(8):2925–2972.

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. Working Paper.

Cassar, L. (2019). Job mission as a substitute for monetary incentives: Benefits and limits. *Management Science*, 65(2):896–912.

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.

Delfgaauw, J., Dur, R., and Souverijn, M. (2020). Team incentives, task assignment, and performance: A field experiment. *The Leadership Quarterly*, 31(3):101241.

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.

Duflo, E., Hanna, R., and Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *American Economic Review*, 102(4):1241–1278.

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, D., Fink, G., and Jack, B. K. (2022). Poor and rational: Decision-making under scarcity. *Journal of Political Economy*, 130(11):2862–2897.

Fehr, E. and Goette, L. (2007). Do workers work more if wages are high? evidence from a randomized field experiment. *American Economic Review*, 97(1):298–317.

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. G. (2013). Teacher incentives and student achievement: Evidence from New York City public schools. *Journal of Labor Economics*, 31(2):373–407.

Fryer, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of Economic Field Experiments*, volume 2, pages 95–322. Elsevier.

Fryer, R. G., Levitt, S. D., List, J., and Sadoff, S. (2022). Enhancing the efficacy of teacher incentives through framing: A field experiment. *American Economic Journal: Economic Policy*, 14(4):269–99.

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.

Gilligan, D. O., Karachiwalla, N., Kasirye, I., Lucas, A. M., and Neal, D. (2022). Educator incentives and educational triage in rural primary schools. *Journal of Human Resources*, 57(1):79–111.

Glazerman, S., McKie, A., and Carey, N. (2009). An evaluation of the teacher advancement program (tap) in chicago: Year one impact report. final report. *Mathematica Policy Research, Inc.*

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., and Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights*, 1(3):291–308.

Goldszmidt, A., List, J. A., Metcalfe, R. D., Muir, I., Smith, V. K., and Wang, J. (2020). The value of time in the united states: Estimates from nationwide natural field experiments. Technical report, National Bureau of Economic Research.

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.

Holmstrom, B. and Milgrom, P. (1991). Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, 7(special_issue):24–52.

Holz, J. E., List, J. A., Zentner, A., Cardoza, M., and Zentner, J. E. (2023). The $ 100 million nudge: Increasing tax compliance of firms using a natural field experiment. *Journal of Public Economics*, 218:104779.

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

Jayaraman, R., Ray, D., and de Véricourt, F. (2016). Anatomy of a contract change. *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.

Kleine, M. (2021). No eureka! incentives hurt creative breakthrough irrespective of the incentives' frame. *Incentives Hurt Creative Breakthrough Irrespective of the Incentives' Frame (May 18, 2021). Max Planck Institute for Innovation & Competition Research Paper*, (21-15).

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., List, J. A., Neckermann, S., and Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, 8(4):183–219.

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. (2020). Non est disputandum de generalizability? a glimpse into the external validity trial. Working Paper.

List, J. A., Livingston, J. A., and Neckermann, S. (2018). Do financial incentives crowd out intrinsic motivation to perform on standardized tests? *Economics of Education Review*, 66:125–136.

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

Loehlin, J. C. and Beaujean, A. A. (2016). *Latent variable models: An introduction to factor, path, and structural equation analysis.* Taylor & Francis.

Loyalka, P., Sylvia, S., Liu, C., Chu, J., and Shi, Y. (2019). Pay by design: Teacher performance pay design and the distribution of student achievement. *Journal of Labor Economics*, 37(3):621–662.

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.

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., and Epstein, S. (2011). *A big apple for educators: New York City's experiment with schoolwide performance bonuses: Final evaluation report.* Rand Corporation.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., and Rajani, R. (2019). Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania. *Quarterly Journal of Economics*, 134(3):1627–1673.

Mbiti, I., Romero, M., and Schipper, Y. (2023). Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania*. *The Economic Journal.*

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.

Miller, G. and Babiarz, K. (2014). Pay-for-performance incentives in low- and middle-income country health programs. In Culyer, A. J., editor, *Encyclopedia of Health Economics*, pages 457–466. Elsevier, San Diego.

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.

Ogundeji, Y. K., Bland, J. M., and Sheldon, T. A. (2016). The effectiveness of payment for performance in health care: a meta-analysis and exploration of variation in outcomes. *Health Policy*, 120(10):1141–1150.

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

Pham, L. D., Nguyen, T. D., and Springer, M. G. (2021). Teacher merit pay: A meta-analysis. *American Educational Research Journal*, 58(3):527–566.

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.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, 37(1):7–63.

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

Salize, H. J., Merkel, S., Reinhard, I., Twardella, D., Mann, K., and Brenner, H. (2009). Cost-effective primary care–based strategies to improve smoking cessation: more value for money. *Archives of Internal Medicine*, 169(3):230–235.

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

Speroni, C., Wellington, A., Burkander, P., Chiang, H., Herrmann, M., and Hallgren, K. (2020). Do educator performance incentives help students? evidence from the teacher incentive fund national evaluation. *Journal of Labor Economics*, 38(3):843–872.

Springer, M. G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J., McCaffrey, D. F., Pepper, M., and Stecher, B. M. (2011). Teacher pay for performance: Experimental evidence from the project on incentives in teaching (point). *Society for Research on Educational Effectiveness*.

Springer, M. G., Pane, J. F., Le, V.-N., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., and Stecher, B. (2012). Team pay for performance: Experimental evidence from the round rock pilot project on team incentives. *Educational Evaluation and Policy Analysis*, 34(4):367–390.

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.

Twardella, D. and Brenner, H. (2007). Effects of practitioner education, practitioner payment and reimbursement of patients' drug costs on smoking cessation in primary care: a cluster randomised trial. *Tobacco Control*, 16(1):15–21.

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

A Supplementary Appendix

A.1 Incentive effects in other field experiments

Table A.1 presents observed effect sizes from a selection of field experiments primarily based on Ogundeji et al. (2016); Bandiera et al. (2021); Fryer et al. (2022), augmented by recently and prominently published studies. The aim of the table is not to provide a comprehensive overview but to illustrate the heterogeneity in effect sizes observed within and across different task categories. The table reports only published field experiments in which some real effort was incentivized with a monetary reward and for which effect sizes in standard deviations were reported in the original study or in one of the three overview studies.

Table A.1: Summary of studies and effect sizes

Reference	Outcome	Incentive	Effect size
Panel A. Retail, production, and service provision			
Delfgaauw et al. (2015)	Average number of products per customer	Elimination tournament with a prize of 1.25% - 6.25% of monthly gross earnings	0.20
Friebel et al. (2017)	Sales	Team bonus of up to €300 per month conditional on pre-existing sales targets	0.33
Fehr and Goette (2007)*	Revenues per four-week period	25 percent increase in commission rate	0.35
Boly et al. (2011)*	Exam-grading accuracy	Monetary penalties for mistakes	0.36
Bandiera et al. (2005)*	Kilograms of fruit picked per hour	Constant piece rate (vs. decreasing piece rate with average productivity of group of workers).	0.86
Hossain and List (2012)*	Inspected units per hour	Roughly 20 percent increase in pay for meeting productivity target	0.90
Panel B. Student performance			
Gneezy et al. (2019)	Mathematics test score (25 questions taken from PISA test), Shanghai	RMB 90 (take away RMB 3.6 for each incorrect answer)	-0.09
Angrist et al. (2009)*	First-year GPA	Monetary bonuses (scholarships) for meeting GPA targets. Higher bonus (up to $5,000) for higher targets.	-0.03
Levitt et al. (2016)	Test score in low stakes tests (low incentive)	$10 incentive for improving their baseline score from the prior testing session	[-0.008; 0.009]
Angrist and Lavy (2009)*	Matriculation-exam performance	Increasing monetary bonuses (up to ca. $1,500) for taking any matriculation test, passing any matriculation test, and completing all matriculation requirements.	0.02
Levitt et al. (2016)	Test score in low stakes tests (high incentive)	$20 incentive for improving their baseline score from the prior testing session	[0.07; 0.15]
Gneezy et al. (2019)	Mathematics test score (25 questions taken from PISA test), US	$25 (take away $1 for each incorrect answer)	[0.24; 0.28]
List et al. (2018)	Test score (knowledge related to ISAT)	$90 if the student improves their score on the test relative to the baseline test	0.30
Panel C. Incentives for teachers			
Fryer (2017)**	Meta-analyses (US experimental studies): Math scores (pooled across years)	–	0.02
Pham et al. (2021)**	Meta-analyses (U.S. experimental and non-experimental studies): Math scores (pooled across years)	–	0.005
Pham et al. (2021)**	Meta-analyses (Non-U.S. experimental and non-experimental studies): Math scores (pooled across years)	–	0.22
Fryer (2013)**	Standardized math performance (pooled across years)	Bonus up to $3,000 for every union-represented teacher	-0.03
Glazerman et al. (2009)**	Standardized math performance	Incentives based on Teacher Advancement Program (TAP, Chicago)	-0.004
Marsh et al. (2011)**	Standardized math performance	Schoolwide Performance Bonuses (New York)	-0.002
Springer et al. (2012)**	Standardized math performance (Round Rock)	$5,400 bonus for teachers in team's who ranked among top third (within grade level)	-0.01
Springer et al. (2011)**	Standardized math performance	bonuses of up to $15,000 per year (for test-score gains on Tennessee Comprehensive Assessment Program (TCAP)	0.003
Mbiti et al. (2019)**	Math scores (Tanzania)	TZS 5,000 (ŨS$ 3) bonus per student passing a standardized test	0.006
Gilligan et al. (2022)**	Math scores (Uganda)	'Pay-for-percentile' incentives (Barlevy and Neal, 2012)	0.02
Duflo et al. (2012)**	Standardized math performance (India)	Incentives for teacher attendance and monitoring	0.02
Behrman et al. (2015)	Student performance (math test, Mexico)	10 and 15 percent of the annual salary of a teacher in a federal high school	[0.01; 0.04]
Speroni et al. (2020)	Student test score (math and reading, year 1-4 pooled)	Bonuses for principals and teacher of an average yearly cost of $100 per student	[0.03; 0.04]
Mbiti et al. (2023)	Student test score (year 1)	'pay for percentile' and 'levels' schemes with multiple (curricular-based) thresholds (3.5% of the annual net salary)	[0.06; 0.22]
Loyalka et al. (2019)**	Math scores (pooled incentive treatments, China)	'Pay-for-percentile' incentives Barlevy and Neal (2012)	0.07
Barrera-Osorio and Raju (2017)**	Student score on government exam (pooled subjects and incentive treatments, Pakistan)	Bonus from 10 – 15% of yearly basic salary	0.08
Fryer et al. (2022)	ThinkLink scores (knowledge related to ISAT)	'pay-for-percentile' (pooled), expected value of reward was $ 4000 (equal to 8 percent of the average teacher salary)	0.10
Muralidharan and Sundararaman (2011)**	Standardized math performance (India)	Average bonus calibrated 3̃% typical annual salary	0.20
Brownback and Sadoff (2020)**	Student score in final exams (multiple postsecondary departments), Indiana community college	Loss-framed incentives of $50 per student who passed an objective, externally-designed course exam	0.20
List et al. (2018)	Student test score (knowledge related to ISAT)	$90 bonus incentive for tutors working with tier two students on ISAT preparation	0.31
Panel D. Health-professionals			
Bardach et al. (2013)***	Cardiovascular risk reduction (US)	Up to $200/patient or $100,000/clinic, paid to clinic	-0.12
Twardella and Brenner (2007)***	Smoking cessation (Germany)	€130 per successful cessation, paid to general practitioners	0.05
Salize et al. (2009)***	Smoking cessation (Germany)	€130 per successful cessation, paid to general practitioners	0.06
An et al. (2008)***	Quitline referrals (US)	$5000 for 50 quitline referrals, paid to clinic	0.06
Basinga et al. (2011)***	Provision and quality of maternal and child health care (Rwanda)	$0.09 - $4.59 per unit of healthcare service, paid to health centre	0.09

Notes: This table presents observed effect sizes from a selection of field experiments primarily based on Bandiera et al. (2021)*, Fryer et al. (2022)** and Ogundeji et al. (2016)*** augmented by recently and prominently published studies. The aim of the table is not to provide a comprehensive overview but to illustrate the heterogeneity in effect sizes observed within and across different task categories. The table reports only published field experiments in which some real effort was incentivized with a monetary reward and for which effect sizes in standard deviations were reported in the original study or in one of the three overview studies.

A.2 Room Fixed Effects for the Natural and Framed Field Experiment

Table A.2: Main treatments probit and GLM regressions including room fixed effects

	Field experiment		Framed field experiment	
	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.6 (4), 5 (4), and A.10 (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.3 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.4 Text of the Invitation to Laboratory Participants

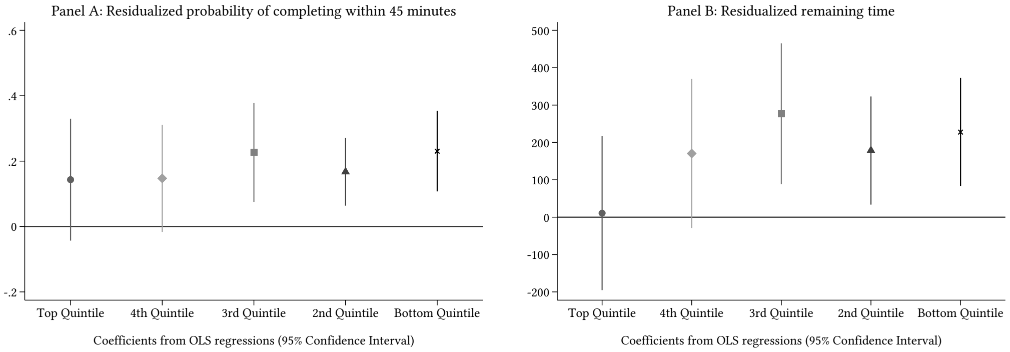
We added the following paragraph to the standard invitation to student participants in the framed field experiment:

"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.5 Additional Analyses for the Field Experiment

A.5.1 Bonus Incentives and Team Characteristics

Table A.3 shows the results from linear probability models estimating a dummy for whether teams complete the task within 45 minutes. Column (1) includes no interactions and uses the same variables and fixed effects as Column (4) in Table 2. The effect of bonus incentives is of a similar magnitude as the average marginal effect in the probit specification. In Columns (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 such as the share of males (2), group size (3), whether the challenge was booked privately (5), or teams' spoken language (6). However, team experience matters for the efficacy of incentives. As shown in Column (4), incentives are substantially more effective for experienced teams, while at the same time the treatment dummy is still statistically significant and large in magnitude. Table A.4 repeats this analysis for teams' remaining times and shows that incentives are more effective for experienced teams (p-value $= 0.10$ for the interaction term in Column (4)). Further, teams with more males have higher remaining times (2), but incentives are less effective for them (in line with the idea of potential ceiling effects). Akin to the analyses regarding



Notes: Panel A shows the effect of incentives on the residualized probabiliy of completing the task within 45 minutes by quintiles. Teams were assigned to quintiles based on their predicted performance using observable team characteristics (and how predictive these were in *Control*). Panel B in the same fashion, shows the effect of incentives on the residualized remaining time by quintiles.

Figure A.1: Incentive effects by quintiles

the probability to finish the task within 45 minutes, other team characteristics do not significantly alter the efficacy of the bonus incentive.

Finally, we shed light on whether the efficacy of incentives differs for teams expected to perform well based on observable characteristics. To do so, we first estimate how observable team characteristics affect their remaining times in *Control.* Based on the obtained coefficients, we then predict for each teams in *Bonus45* and *Control* their performance, and sort all teams into the respective quintiles. We build the residualized completion probability and remaining time by subtracting the predicted performance from actual performance. In a second step, we estimate the treatment effect for the residualized probability to solve the task within 45 minutes (Panel A of Figure A.1) and teams' remaining times (Panel B of Figure A.1) for each quintile.

Both panels show that incentives do not statistically significantly improve outcomes for teams in the top quintile, who may already be exerting a lot of effort and thus have less scope for improvement. Notably, incentives are effective for all other quintiles. For these, we observe strong and statistically significant effects.[59]

[59] p-values for the 4th quintile are $p = 0.093$ for the residualized remaining time, and $p = 0.077$ for the residualized completion probability.

Table A.3: 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 2. Robust standard errors clustered at the day level reported in parentheses, and *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A.4: GLM regressions: remaining time

	GLM: Remaining time in seconds					
	(1)	(2)	(3)	(4)	(5)	(6)
Bonus45	0.257**	0.569***	0.612	0.154	0.256*	0.276**
	(0.116)	(0.209)	(0.482)	(0.127)	(0.155)	(0.125)
Share males	0.513***	0.867***	0.513***	0.509***	0.513***	0.507***
	(0.134)	(0.159)	(0.132)	(0.130)	(0.134)	(0.137)
Group Size	0.286***	0.287***	0.327***	0.288***	0.286***	0.287***
	(0.048)	(0.047)	(0.052)	(0.048)	(0.048)	(0.048)
Experience	0.336***	0.343***	0.334***	0.186	0.336***	0.340***
	(0.086)	(0.086)	(0.086)	(0.121)	(0.087)	(0.084)
Private	0.197**	0.195*	0.196**	0.188**	0.195	0.198**
	(0.098)	(0.104)	(0.098)	(0.095)	(0.162)	(0.098)
English-speaking	-0.333 (0.240)	-0.352 (0.235)	-0.333 (0.236)	-0.347 (0.237)	-0.334 (0.241)	-0.160 (0.201)
Bonus45 ...						
... × Share males		-0.562**				
		(0.256)				
... × Group Size			-0.072			
			(0.086)			
... × Experience				0.244		
				(0.148)		
... × Private					0.003	
					(0.177)	
... × English-speaking						-0.281 (0.460)
Constant	4.136***	3.929***	3.942***	4.162***	4.137***	4.073***
	(0.387)	(0.356)	(0.369)	(0.384)	(0.401)	(0.373)
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 generalized linear model with log link. Dependent variable: Remaining time in seconds. All models include 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.5.2 Probability of Completing the Task in 45 Minutes (Field Experiment)

Table A.5 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.5: 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)
Bonus45	0.150***	0.151***	0.183***	0.163***
	(0.026)	(0.024)	(0.027)	(0.045)
Fraction of control teams completing the task in less than 45 min	0.11	0.11	0.11	0.11
Control Variables	No	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes
Week Fixed Effects	No	No	No	Yes
Observations	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 indicator. 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.6). 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.

Table A.6: GLM regressions: Remaining time

	GLM: Remaining time in seconds			
	(1)	(2)	(3)	(4)
Bonus45	0.432***	0.447***	0.406***	0.257**
	(0.088)	(0.096)	(0.094)	(0.116)
Constant	5.842***	4.041***	4.251***	3.803***
	(0.082)	(0.393)	(0.359)	(0.403)
Control Variables	No	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes
Week Fixed Effects	No	No	No	Yes
Observations	487	487	487	487

Notes: Coefficients from a generalized linear model regression with a log link of the remaining time on our treatment indicator. Control variables, staff, and week fixed effects as in Table 2. Robust standard errors clustered at the day level reported in parentheses, and *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A.7: 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 *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

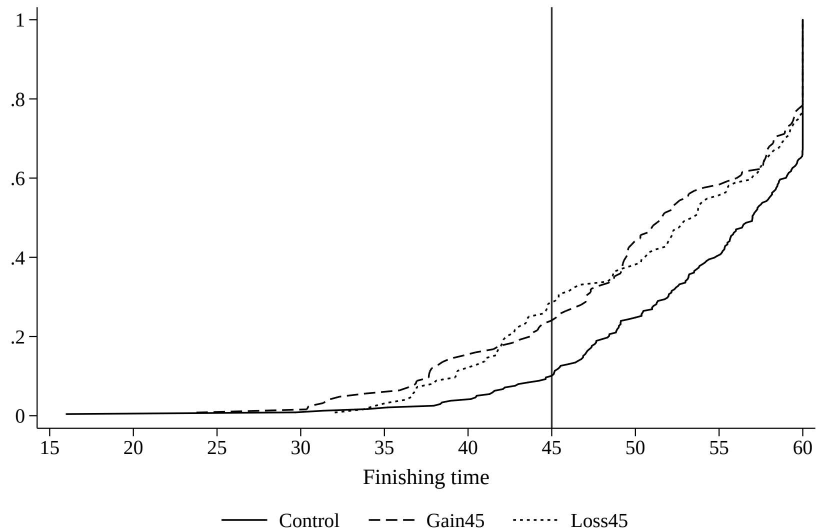
Analogously to the probit regressions reported in Table 4, we also run GLM specifications with the remaining time as the dependent variable (Table A.7) for the full set of incentive treatments. This confirms our findings that incentives that include rewards increase performance whereas only mentioning the reference performance does not.

A.5.4 Framing of Bonus Incentives (Field Experiment)

As explained in Section 2, 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.[60] Figure A.2 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 times for the two framings come from the same underlying distribution (p-value $= 0.70$). Also, the fraction 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 those 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 A.8 summarizes these different performance measures and Table A.17 in Section A.10 highlights that the observed incentive effect is robust to controlling for multiple hypotheses testing using procedures recommended in List et al. (2019).

[60]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.



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 A.2: Finishing times in *Gain45*, *Loss45*, and *Control* in the field experiment

Table A.8: 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 Appendix Table A.17 for MHT adjusted p-values according to List et al. (2019).

A.6 Additional Analyses for the Framed Field Experiment

A.6.1 Overview of Performance Across Treatments (Framed Field Experiment)

Table A.9 provides an overview of the fraction of teams finishing the task within 45 (60) minutes as well as the remaining times across treatments.

Table A.9: 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.

A.6.2 Regression Analysis for Remaining Time as Dependent Variable (Framed Field Experiment)

Table A.10 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 $= 0.88$).

A.6.3 Probability of Completing the Task in 45 Minutes (Framed Field Experiment)

Table A.11 reports the results for the regression columns (1) to (5) from Table 5 excluding those weeks where we do not observe variation in the outcome variable. This confirms our previous findings.

Table A.10: 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 5. Robust standard errors clustered at the session level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.11: 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)
Bonus45	0.107*	0.097*	0.104**	0.111**
	(0.055)	(0.054)	(0.052)	(0.051)
Fraction of control teams completing the task in less than 45 min	0.06	0.06	0.06	0.06
Control Variables	No	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes
Week Fixed Effects	No	No	No	Yes
Observations	191	191	191	191

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 5. 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.6.4 Factor Analyses (Questionnaires in Framed Field Experiment)

Table A.12: Factor analyses

	Control	Bonus45	p-values
Questionnaire 1			
Factor 1 (*Collaboration*)	-0.0015	0.0007	0.7631
Factor 2 (*Team Cohesion*)	-0.0958	0.0469	0.0715
Factor 3 (*Dominance*)	-0.0834	0.0408	0.1056
Factor 4 (*Documentation*)	0.0853	-0.0417	0.0155
Factor 5 (*Intensity*)	-0.1478	0.0723	0.0066
Observations	264	540	804
Questionnaire 2			
Factor 1 (*Cooperative*)	0.0311	-0.0146	0.3406
Factor 2 (*Leadership*)	-0.2244	0.1054	0.0038
Factor 3 (*Struggling*)	0.0960	-0.0451	0.2572
Observations	117	249	366

Notes: This table reports means of factors based on factor analyses of two questionnaires as part of the framed field experiment. For Questionnaire 1 five factors survived the factor analyses, while three factors survived the analyses for Questionnaire 2. Items from Questionnaire 1 that load heavy on Factor 1 are: (5) (6), (7), (9), (13), (15), and (18). The items loading heavy on Factor 2 are: (8), (10), (12), (16), (17), and (19). The items loading heavy on Factor 3 are: (2) and (4). The items loading heavy on Factor 4 are: (3), (11). The items loading heavy on Factor 5 are: (1), (14). Items from Questionnaire 2 that load high on Factor 1 are: (1, negatively), (5), (7), (8), (10), (11), and (12). Items that load high on Factor 2 are: (2) and (6). Items that load high on Factor 3 are: (3), (4), and (9). Numbers in parentheses refer to the questions in Table 8. The last column contains p-values from Mann-Whitney tests.

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

Table A.13: Ordered probit regressions: Number of hints requested

	Ordered probit: Number of hints requested							
	Field experiment				Framed field experiment			
	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) are 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

Table A.14: 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.

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.14 we 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.15).

Table A.15: 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 5. 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. However, 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.16 shows that both below median risk-taking and the interaction term do not significantly affect hint-taking behavior. Columns (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.16: OLS regressions: Number of hints requested

	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) are 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 Multiple Hypotheses Testing (adjusted p-values)

A.10.1 Field Experiment

Table A.17 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 differ significantly at the ten percent level when performing twelve tests simultaneously. Outcomes in *Gain45* and *Loss45* treatments do not differ.

Table A.18 relates to Table 4 and presents adjusted p-values by simultaneously testing for differences in multiple outcomes and treatments (*Bonus45*, *Bonus60*, and *Reference*

Table A.17: Field experiment - MHT adjusted p-values according to List et al. (2019) (referring to Table A.8)

Outcome	*Control* vs. *Bonus45*	*Control* vs. *Gain45*	*Control* 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.18: Field experiment - MHT adjusted p-values according to List et al. (2019)

Outcome	*Control* vs. *Bonus45*	*Control* vs. *Bonus60*	*Control* 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.

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.

A.10.2 Framed Field Experiment

Table A.19 refers to Table A.9 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

Table A.19: Framed Field experiment - MHT adjusted p-values according to List et al. (2019) (referring to Table A.9)

Outcome	*Control* vs. *Bonus45*	*Control* vs. *Gain45*	*Control* 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	0.4430
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.

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.11 Additional leadership analyses

Table A.20: GLM: Leadership: Remaining time

	Remaining Time in Seconds			
	(1)	(2)	(3)	(4)
Leadership	0.542***	0.550***	0.544***	0.598***
	(0.167)	(0.164)	(0.191)	(0.190)
Mean in Control	191.1	191.1	191.1	191.1
Controls	No	Yes	Yes	Yes
Staff FE	No	No	Yes	Yes
Week FE	No	No	No	Yes
Observations	281	281	281	281

Notes: The table displays coefficients from generalized linear model regressions with a log link of remaining time on our *Leadership* indicator (with *Control-L* as base category). Each column indicates whether team controls (group size, share of male participants, experience with escape games, median age, language spoken, private versus team-building events, actively taken walkie-talkie), staff, and week fixed effects are included. Standard errors in parentheses are clustered at the daily level, with significance levels * = $p < 0.10$, ** = $p < 0.05$, and *** = $p < 0.01$.

Table A.21: Team performance (completion and remaining time with room fixed effects)

	Completed within 60 Minutes				Remaining Time in Seconds			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Leadership	0.137***	0.137***	0.125**	0.105**	0.378***	0.354***	0.298**	0.321**
	(0.045)	(0.047)	(0.058)	(0.043)	(0.144)	(0.126)	(0.143)	(0.161)
Mean in Control	0.442	0.442	0.442	0.442	191.1	191.1	191.1	191.1
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Staff FE	No	No	Yes	Yes	No	No	Yes	Yes
Week FE	No	No	No	Yes	No	No	No	Yes
Observations	281	281	281	281	281	281	281	281

Notes: The table displays average marginal effects from Probit regressions of whether a team completed the task within 60 minutes (Columns (1)–(4)) and coefficients from generalized linear model regressions with a log link of remaining time (Columns (5)–(8)) on our *Leadership* indicator (with *Control-L* as base category). All columns include room fixed effects. Each column indicates whether team controls (group size, share of male participants, experience with escape games, median age, language spoken, private versus team-building events, actively taken walkie-talkie), staff, and week fixed effects are included. Standard errors in parentheses are clustered at the daily level, with significance levels * = $p < 0.10$, ** = $p < 0.05$, and *** = $p < 0.01$.

A.12 Incentives and Effort Dimensions

A.12.1 Expert Survey

In addition to highlighting (the demand for) leadership as a central mechanism of how incentives improve team performance, we also explore which effort dimensions may be affected by incentives in non-routine team tasks. Based on numerous comments in seminars, workshops, and conference presentations, we compiled a list of ten potentially important effort dimensions (see Table A.22) 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.[61] 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 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

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

Table A.22: Survey results

Statement *With incentives, ...*	Average Rank Experts	 ESA	# of Wins in Pair-wise Comparisons 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 respondents deem this dimension).

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.[62] 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.

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 fill in the survey carefully. We also removed those who did not rank all dimensions, leaving us with 28 and 65 responses respectively.

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

Table A.22 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 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"*.

A.12.2 Additional laboratory experiment: Description

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 had ranked as most important.

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 sub-task (*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*.

As participants arrived at the laboratory, they were randomized into teams of three and each team was guided to a separate room to perform the task (with treatments being randomized across these rooms as well). In each room, one experimenter welcomed the participants and explained the general procedures, before each participant underwent a cognitive skill test (Raven's progressive matrices) on a computer tablet at a separate workstation. After completion of the test, each participant received their own test score as private information but no participant was informed about their team members' performance in the test. Then, all three participants were guided to stand around a large table in the middle of the room, to perform the board game escape challenge.

The board game escape challenge was framed as a secret mission, in which participants needed 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). Participants were guided by a tablet computer placed in the middle of the table. The tablet displayed the time left to solve the escape challenge and served to electronically record task solutions entered by the team. Further, the tablet displayed 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 original board game our team challenge is based on). To take notes, each participant received a pen and a paper, and each team member was equipped with an identical decoding sheet. Further, each team member received 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 that they were ready to commence 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’, a solution 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.[63]

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.[64] They could then use the solution 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.[65] 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 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.[66]

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

[63]Each time participants failed to enter the correct code, 3 minutes were subtracted from the available time.

[64]Each time participants chose the wrong flag, 3 minutes were subtracted from the available time.

[65]If participants failed to enter the correct key code, 1 minute was subtracted from the available time.

[66]Each time participants chose the wrong enemy country, 3 minutes were subtracted from the available time.

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 difference in difference analysis.[67]

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 also not share the information and use the materials provided to work on the part's solution. By comparing 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, 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.[68]

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 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 the respective materials and instructions.

[67]As we do not find that incentives significantly affect the extent of communication reported by our participants, we refrain from including such a differences in difference analysis 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).

[68]Each time participants entered a wrong key code, 1 minute was subtracted from the available time.

The individual task required to sort 8 picture cards (with pictures on both sides) into a 2 $\times$ 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.[69] 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.[70]

After participants entered the correct four keys (or if the 60 minutes expired, whichever occurred first), the task ended and participants filled in a short survey, including a question on the extent of communication within the team, 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 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*) would have allowed 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 observed similar finishing times and variances as in the framed field experiments, we could 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 was expected to be lower for binary outcomes such as finishing within 60 or 45 minutes. Using a χ^2-test, we could identify effect sizes

[69]Each time the participant entered a wrong solution, 1 minute was subtracted from the available time.
[70]Each time participants entered a wrong key code, 1 minute was subtracted from the available time.

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 needed 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 removed 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 below 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 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 substantial treatment effects for these outcome variables, we refrain from reporting additional robustness checks (see also our pre-registration for the additional laboratory experiment).

A.12.3 Additional laboratory experiment: Results

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).[71] 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 solving the task within the incentive target of 45 minutes (*BGIncentive45*: 7 percent, *BGControl*: 2 percent). Due to substantial noise in the data 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 also more 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 $= 0.40$).

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

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*).

A.13 Surveys with HR Experts

To quantify a reasonable prior of the effectiveness of incentives in non-routine analytical team tasks, in February 2023, we surveyed 400 participants from a pool of HR experts, who were responsible for making hiring decisions in their jobs.[72] The sample was provided by survey provider Cint. To compare expectations about non-routine and routine tasks, we randomly assigned about half of these experts (n=197) to a condition in which expectations about non-routine analytical team tasks in general were elicited. We explained to these participants that non-routine analytical tasks require problem solving, intuition, or creativity and are often found in occupations that encompass executive or managerial functions, technical, or creative occupations (for example, lawyers, medical and engineering professions, designers, and managers), while routine tasks were explained as tasks that can also be specified to be performed by a machine and are typically found in occupations with medium educational requirements (for example: accounting, secretarial tasks, industrial production, monitoring). To study how expectations about non-routine tasks in general differ from expectations in the context of escape challenges, we assigned 99 HR experts to a condition in which they were explicitly asked about the effectiveness of bonus incentives in escape challenges (i.e., in addition to the task description mentioned above, they learned about the specifics of the setting). Finally, to elicit informed expectations regarding our particular setting, we provided 104 HR experts (out of the 203 who were assigned to state expectations about escape games) with team' average remaining times in our Control condition (these teams had on average 6 minutes

[72]The survey was pre-registered as AsPredicted (#122060), `https://aspredicted.org/1SW29C`.

remaining) before eliciting experts' expectations about the incentive effect in the escape challenge.

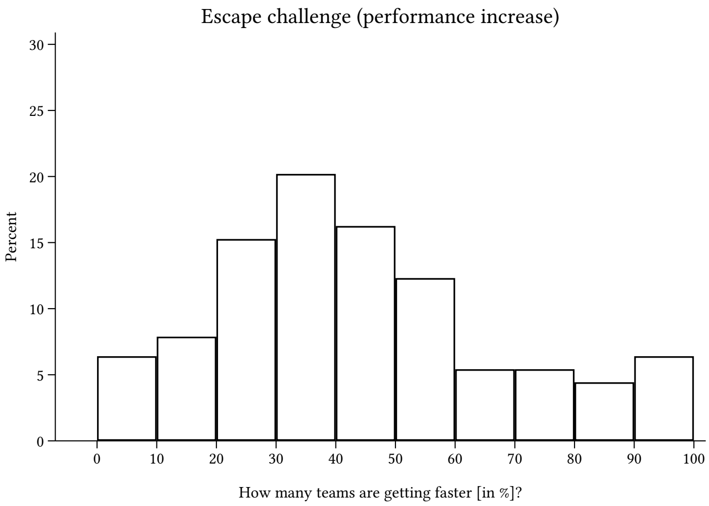
In all conditions, HR Experts had to indicate how many out of one hundred teams would i) become faster, ii) slower, and iii) do neither, once they receive the opportunity to earn a bonus. Additionally, as participants could have believed that improving teams became substantially faster, whereas declining teams only moderately slower, we also asked for the number of minutes teams would be expected to become faster/slower (conditional on becoming faster/slower), which allows us to calculate the average expected change in performance (in minutes). Translated survey instructions can be found below.

The results from the HR expert survey are summarized in Table A.23. Regarding the 197 HR experts who formed expectations about abstract tasks, we find that the average expected improvement due to incentives amounts to 3.22 minutes in non-routine tasks and to 4.13 minutes in routine tasks and differs statistically significantly (Wilcoxon signed-rank test, p= 0.013). These experts are also more optimistic regarding the fraction of teams that improve with incentives (Wilcoxon signed-rank test, p= 0.018). Further, we find that experts' expectations about performance improvements in escape challenges are similar to expectations about non-routine tasks more generally (Mann-Whitney test, abstract vs. no info: $p = 0.993$, abstract vs. info: $p = 0.348$, pooled: $p = 0.556$). Finally, the survey revealed substantial heterogeneity in expectations within and across experts. On average, experts expect performance improvements for 39–42 percent of teams in the escape challenges. While the median HR expert expects 40 out of 100 teams to improve when facing incentives, 20% of them believe that between 0-20 teams will improve, while another 20% believe that 60-100 teams will improve (see also Figure A.3.)

Table A.23: Expected effect sizes

	Fraction of teams faster in %	slower in %	same in %	Improvement in minutes
Abstract (n=197)				
Non-routine task	41.37	21.48	37.15	3.22
Routine task	44.55	20.23	35.22	4.13
Escape challenge (n=203)				
Escape (no info, n=99)	42.05	22.18	35.77	3.77
Escape (info, n=104)	38.80	24.42	36.78	1.97
Escape (pooled)	40.38	23.33	36.29	2.84

Notes: This table reports means of survey answers on how many teams are getting faster, slower or are not affected by a bonus incentive. Additionally, the overall expected improvement (average reduction in finishing times) is reported.



Notes: The figure shows histograms of survey answers on how many teams they expect to become faster in an escape challenge, when there is a bonus incentive in place.

Figure A.3: Expected performance increase

Translated instructions

(text in square brackets only visible to participants in respective treatment condition)

Welcome!

For this survey, we want to collect your assessments of the effects of financial incentives in various team tasks. To this end, we will first provide some definitions:

Routine tasks:
Any type of task that can be specified to be performed by a machine (for example: adding multiple numbers). Routine tasks are typical of many occupations with intermediate educational requirements, for example, accounting, secretarial tasks, industrial production, or supervision.

Non-routine tasks:
Any type of task that requires problem solving, intuition, persuasion, or creativity. These tasks are often found in occupations involving managerial, technical, or creative tasks, for example, lawyers, medical and engineering occupations, designers, and managers.

[**Abstract**: For the following questions, imagine a non-routine work environment in which workers in a team must complete a series of complex tasks. All tasks must be successfully completed within one hour (= 60 minutes). There is also a possibility that not all tasks will be successfully completed after the time has elapsed].

[**Escape**: For the following questions, imagine an Escape Game as an example of a non-routine task. In Escape Games, teams must solve a series of complex tasks to escape from a room. To do this, teams must find various clues, combine information, and think around corners. All tasks must be successfully completed within one hour (= 60 minutes). There is also a possibility that not all tasks will be successfully completed after the time has elapsed].

In addition to the usual reward, there is a consideration to introduce a bonus for the whole team, which the team will receive if the tasks are successfully completed after 45 minutes already.

[**Escape Info**: Assume that teams that are not offered a bonus will, on average, have successfully completed all tasks about 6 minutes before time expires].

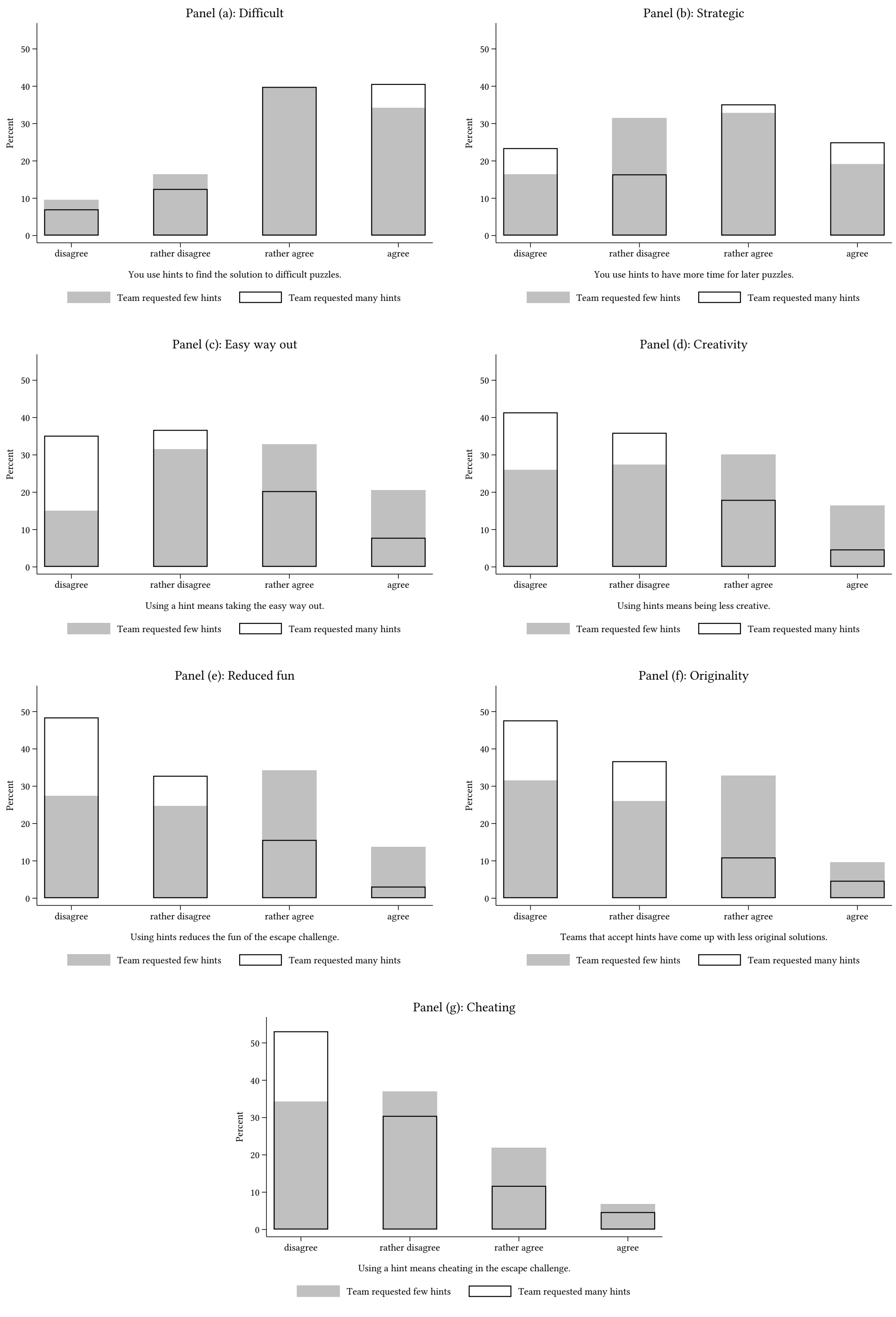
A.14 Additional Customer Surveys on Goals and Hint Taking

To identify how teams' goals are potentially shifted when teams face incentives as well as how teams perceive hint taking, we ran additional surveys with 201 customers performing the team challenge at ETR Munich in January 2023.[73]

Prior to their participation in the escape challenge, survey participants were asked to rank eight potential goals they may pursue in the challenge from most (rank 1) to least important (rank 8). Half of the participants were asked to rank goals for a hypothetical scenario in which they had the opportunity to win a team bonus of 50 Euro if they completed the task in 45 minutes ('Bonus' condition, $n = 100$). The other half was randomly assigned to a 'No bonus' condition ($n = 101$), i.e., they ranked the goals without any bonus being mentioned. Table A.24 summarizes our findings. As can be seen, teams care about being successful in a challenging task, uphold a good atmosphere within the team, and get out of the room as quickly as possible. Teams also consider taking no hints as a potential goal whereas getting to know team members, competition within teams (being better than one's team members) or staying in the room for long are considered the least important. Interestingly, bonus incentives offered for performance do not strongly affect how goals are ranked. The only statistically significant difference exists for the goal of solving more tasks than a team member (Mann-Whitney test, $p = 0.095$), which seems to be more important when there is no incentive scheme in place. Furthermore, while there is no statistically significant difference between the two conditions for the goals of creating a good atmosphere in the team and facing a challenge, the ordering slightly differs.

After their participation in the escape challenge, survey participants had to evaluate by how much they agree with seven statements about hint taking. Our findings are summarized in Figure A.4. To capture potential image concerns, Figure A.4 shows histograms of responses split by the number of hints taken by these teams. We define teams with less than three hints as those taking few hints, and teams with three or more hints as teams taking many hints. Both, teams taking many hints and teams taking few hints agree that hints are used to find the solution to difficult puzzles (χ^2 test: p-value = 0.71) and they only have small disagreements over hint usage being done to have more time for later puzzles (χ^2 test: p-value = 0.082). Teams that take many hints tend to perceive hint

[73]The additional survey was and pre-registered at AsPredicted (#117067), `https://aspredicted.org/ZKKNCS`.



Notes: The figure shows histograms of survey answers on the perceptions about hint taking for teams that took many hints (three or more hints) and team that took few hints (two or less hints). For each of the seven statements, subjects had to evaluate whether they disagree or agree with the respective statement on a 4-point Likert-scale.

Figure A.4: Perceptions about hint taking

Table A.24: Goals of participating in an escape challenge

Statement	Average Rank		Wilcoxon rank-sum	# of Wins in Pairwise Comparisons	
I want to...	Bonus	No bonus	test (p)	Bonus	No bonus
... *achieve success together.*	2.66	2.56	0.71	7	7
... *create a good atmosphere in the team.*	3.18	3.51	0.24	6	5
... *face a challenge.*	3.38	3.42	0.86	5	6
... *get out of the room as quickly as possible.*	3.98	4.32	0.28	4	4
... *take no hints.*	4.56	4.73	0.51	3	3
... *get to know my team members better.*	5.52	5.33	0.51	2	2
... *solve more tasks than my team members.*	6.16	5.61	0.10	1	1
... *stay in the room as long as possible.*	6.56	6.51	0.84	0	0
Observations	100	101	201	100	101

Notes: This table reports how customers of ETR rank different goals of participating in an escape challenge. Customers in *Bonus* were asked to rank these goals when a bonus incentive is in place. Average rank reports the average rank assigned to a statement (from 1 to 8) across all respondents within the respective sample (i.e., the lower the 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).

taking less often as the easy way out (χ^2 test: p-value < 0.01), but absolute differences are again small. Clearly, teams using few hints are more likely to agree that hint taking reduces fun (χ^2 test: p-value < 0.01), is less creative (χ^2 test: p-value < 0.01), reduces originality (χ^2 test: p-value < 0.01), and can be considered cheating (χ^2 test: p-value = 0.05). As such, it becomes clear that teams may refrain from hint taking if they have an intrinsic motivation to explore on their own or perceive taking hints as negative signals about their own creativity or integrity.