The Effect of Incentives in Nonroutine Analytical Team Tasks

Florian Englmaier

Ludwig Maximilian University Munich

Stefan Grimm

Ludwig Maximilian University Munich

Dominik Grothe

Ludwig Maximilian University Munich

David Schindler

Tilburg University

Simeon Schudy

Ulm University

Despite the prevalence of nonroutine analytical team tasks in modern economies, little is understood regarding how incentives influence performance in these tasks. In a series of field experiments involving more than 5,000 participants, we investigate how incentives alter

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, Anastasia Danilov, Wouter Dessein, Robert Dur, Florian Ederer, Constança Esteves-Sorenson, Armin Falk, Urs Fischbacher, Guido Friebel, Svenja Friess, Uri Gneezy,

Electronically published June 5, 2024

Journal of Political Economy, volume 132, number 8, August 2024.

https://doi.org/10.1086/729443

behavior in teams working on such a task. We document a positive effect of bonus incentives on performance, even among teams with strong intrinsic motivation. Bonuses also transform team organization by enhancing the demand for leadership. Exogenously increasing teams' demand for leadership results in performance improvements comparable to those seen with bonus incentives, rendering it as a likely mediator of incentive effects.

I. Introduction

Until the 1970s, a major share of the workforce performed predominantly manual and repetitive routine tasks with little need for coordination in teams. Since then, the work environment has rapidly changed. Nowadays, work is frequently organized in teams (see, e.g., Bandiera, Barankay, and Rasul 2013), and a large share of the workforce performs tasks that require a greater amount of cognitive effort compared with physical labor.

Examples include teams of information technology (IT) professionals, specialist doctors, and management consultants. These teams often face a series of novel and complex problems and need to gather, evaluate, and recombine information to succeed, frequently in a limited amount of time. Autor, Levy, and Murnane (2003) analyze task input in the US economy using four broad task categories: routine manual tasks (e.g., sorting or repetitive assembly), routine analytical and interactive tasks (e.g., repetitive customer service), nonroutine manual tasks (e.g., truck driving), and nonroutine analytical and interpersonal tasks (e.g., forming and testing hypotheses). They document a strong increase in the latter category

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, and Roberto Weber as well as seminar participants at University of Augsburg, University of Barcelona, University of Bonn, Central European University, Columbia University, Heidelberg University, Johns Hopkins University, Karlsruhe Institute of Technology, University of Lausanne, Ludwig Maximilian University (LMU) Munich, National Bureau of Economic Research OrgEc Meeting, University of Regensburg, Tilburg University, Wharton School of the University of Pennsylvania, and numerous conferences for very helpful comments. The helpful comments and suggestions from four referees and the editor substantially improved this manuscript. 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 Wicker, Nicolas Wuthenow, and Aloysius Widmann for excellent research assistance. David Schindler gratefully acknowledges funding by the Dutch Research Council (NWO) under the Talent Scheme (file VI.Veni.211E.002). Stefan Grimm acknowledges funding by the German Research Foundation (DFG) through Graduiertenkolleg 1928. Financial support by the DFG through Collaborative Research Center Transregio 190 (project 280092119) is also gratefully acknowledged. This study was approved by the Department of Economics Institutional Review Board at LMU Munich (project 2015-11). This paper was edited by John List.

between 1970 and 2000. Autor and Price (2013) reaffirm the importance of these tasks in 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., Erev, Bornstein, and Galili 1993; Lazear 2000; Shearer 2004; Bandiera, Barankay, and Rasul 2005, 2013; Hossain and List 2012; Delfgaauw et al. 2015; Jayaraman, Ray, and de Véricourt 2016; Englmaier, Roider, and Sunde 2017; Friebel et al. 2017)—evidence on the effects of bonus incentives is scarce for nonroutine analytical tasks where teams collaboratively solve complex problems.[1]

The efficacy of incentives may substantially differ in nonroutine analytical team tasks for several reasons. First, they are often performed by people who enjoy their challenging nature and are intrinsically motivated (see, e.g., Friebel and Giannetti 2009; Delfgaauw and Dur 2010; Autor and Handel 2013).[2] In turn, extrinsic incentives could negatively affect team performance by crowding out workers' intrinsic motivation (e.g., Deci, Koestner, and Ryan 1999; Hennessey and Amabile 2010; Eckartz, Kirchkamp, and Schunk 2012; Gerhart and Fang 2015). Bénabou and Tirole (2003) provide a theoretical framework formalizing arguments for crowding out on the basis of the idea that incentives may alter workers'

[1] This study focuses on performance-related bonus payments that 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% of which use short-term incentives and >50% use long-term ones), with, on average, 76% of firms using annual incentive plans. However, the use of different annual incentive pay components varies substantially across firms and levels, rendering the question of whether bonus incentives work in nonroutine tasks crucial from a practitioner's perspective. For a more general discussion on the use of performance-related bonus payments as part of compensation in firms, see also Moynahan (1980), Churchill, Ford, and Walker (1993), Prendergast (1999), Lazear (2000), Oyer (2000), and Lazear and Oyer (2013). For theoretical motivations to use simple binary payment schemes, see, e.g., Fehr, Klein, and Schmidt (2007), Herweg, Müller, and Weinschenk (2010), Larkin and Leider (2012), and Ulbricht (2016).

[2] Intrinsic motivation may stem from direct task utility (and thus reflect lower levels of 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 from greater goals attached to the activity (such as job mission; see, e.g., Cassar 2019). We do not consider greater goals or job missions to be necessary in all nonroutine team tasks. However, we believe that both direct task utility and benefits beyond the production outcome are often relevant in nonroutine 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).

perception of the task or their own ability. For example, they 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 nonroutine tasks are generally multidimensional, incentives may lead to crowding out because of 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., McCullers 1978; McGraw 1978; Amabile 1996; Azoulay, Graff Zivin, and Manso 2011; Ederer and Manso 2013).

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 nonroutine 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 nonroutine 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 (because of intrinsic motivation, status, recognition, or career concerns).

Third, incentives may be less effective in team settings, as free riding could be present. The output produced by some workers can be misattributed to the work of others, and, additionally, team incentives reward the overall team output instead of individual contributions (Holmstrom 1982). Fourth, salient incentives may also alter team organization (Englmaier, Roider, and Sunde 2017). Particularly in nonroutine tasks, incentives may create a demand for efficient leadership that enables teams to solve complex problems in a more coordinated manner. The variety of reasons for why incentives may work differently in nonroutine analytical tasks is mirrored in substantial heterogeneity in experts' expectations about the efficacy of incentives, underscoring the need for clean empirical evidence on how incentives alter behavior in teams collaboratively performing nonroutine analytical tasks.[4]

This study exploits a unique field setting to measure the effects of bonus incentives for behavior in teams collaboratively performing a nonroutine

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

[4] For instance, we document in an additional survey with human resource (HR) experts that the range of predictions of incentive efficacy varies strongly. While the median HR expert expects 40 out of 100 teams to improve when facing incentives, 20% of them believe that 0–20 teams will improve, while another 20% believe that 60–100 teams will improve (see fig. A.8 for the full distribution and app. sec. A.16 for more details on the survey; the appendix, including figs. A.1–A.8, is available online).

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 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 nonroutine analytical and interactive team tasks: teams face a series of complex and novel problems, need to collect and recombine information, and must solve analytical and cognitively demanding quests that require thinking outside the box. The task is also interactive since members of each team have to collaborate with each other, discuss possible actions, and develop ideas jointly. At the same time, real-life escape games allow for an objective measurement of joint team performance (time spent until completion) as well as for exogenous variation in incentives for a large number of teams.

Our setting is particularly flexible, allowing us to vary the incentive structure for over 700 teams (3,308 participants) under otherwise equal conditions and to replicate the main findings in a second distinct sample of presumably less intrinsically motivated teams (268 teams, 804 participants). Further, it enables us to identify potential mechanisms behind the effects of bonus incentives by running an additional field experiment (281 teams, 1,273 participants), hence substantially advancing the literature on the effects of incentives in collaboratively solved nonroutine team tasks.

To identify the causal effects of incentives on behavior, we first conducted a series of field experiments with strongly intrinsically motivated teams (which were regular participants in escape games at ExitTheRoom [ETR], a firm we partnered with) who were unaware of taking part in an experiment.[5] We implemented a between-subject design, in which teams were randomly assigned to either a treatment or a control condition. For the main treatment, we offered a team bonus (of approximately €10 per participant) if the team completed the task within 45 minutes (the regular prespecified 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 are more than twice as likely to complete the task within 45 minutes. Moreover, in line with the idea that nonroutine tasks feature an important noisy component in how effort translates into performance, bonus incentives not only induce a

[5] Harrison and List (2004) classify this approach as a natural field experiment. The study was approved by the Department of Economics Institutional Review Board 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.

local effect around the threshold for receiving the bonus but also improve performance over a significant part of the distribution of finishing times.[6]

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 combines the bonus payment with a rather ambitious performance threshold (45 minutes), which could be interpreted as a positive signal about workers' ability. Further, this ambitious performance threshold itself could cause performance improvements (independent of the bonus incentive).

To test for such countervailing effects, we implement two additional treatment conditions. We first combine the bonus with a less ambitious performance threshold (60 minutes) and thus provide additional room for crowding out due to incentives. The second condition provides the ambitious threshold (45 minutes) as a reference point, signaling excellent performance but no monetary reward. The 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.[7] Hence, it is unlikely that the existence

[6] Many nonroutine tasks may feature a noisy production function or (low) effort elasticity, which may, in turn, reduce bunching around bonus thresholds or performance goals. In contrast, bunching can occur in routine tasks, where the relationship between effort provision and outcomes is more deterministic and oftentimes precise, and (real time) feedback about performance is available (see, e.g., Hossain and List 2012; Allen et al. 2017; Kuhn and Yu 2021). However, routine tasks may also not exhibit bunching resulting from strategic responses to incentives, e.g., when feedback is noisy, provided only on an aggregate level, or with delay. For instance, Friebel et al. (2017) do not find differences in the distributions of the percentage of sales (as a percentage of the target) between their treatment and control teams, indicating that their incentive condition did not result in bunching (we thank the authors for reporting these additional results to us). The latter aspects as well as a potential lag of continuous outcome variables may explain why several other field experiments related to bonus incentives in routine tasks (see table A.1; tables A.1–A.24 are available online) do not report bunching.

[7] The latter findings also complement recent research on nonmonetary 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., Azmat and Iriberri 2010; Barankay 2010, 2012; Blanes i Vidal and Nossol 2011). Our finding, however, does not rule out that salient performance goals may further increase team performance, as observed, e.g., in laboratory (Corgnet, Gómez-Miñambres, and Hernán-Gonzalez 2015) and field experiments (Gosnell, List, and Metcalfe 2020).

of the bonus incentive strongly crowded out teams' intrinsic motivation to solve the task quickly.[8]

Second, in the spirit of List (2003, 2004a, 2004b, 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 likely less intrinsically motivated, as they did not self-select into the task.[9] We find that despite potentially lower intrinsic motivation, bonus incentives similarly improve performance. Akin to the results from the field experiment, incentives more than double the fraction of teams that manage 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, we find that incentives do not significantly reduce the willingness of teams to explore original solutions among likely more intrinsically motivated customer teams, who self-select into the task. However, we observe an increase in hint taking due to incentives among the presumably 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.[10]

As 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 customer teams. This suggests that awareness of how effort translates into performance enhances the positive incentive effect.

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

[8] 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 a bonus is offered (see also table A.23).

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

[10] 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 sec. III.C.3; fig. A.7).

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 this experiment, we exogenously varied the demand for leadership by nudging (or not nudging) teams to pick a leader. The experiment reveals a substantial positive effect of leadership demand on team performance. The findings are 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; rather, they appear to provide the impetus for teams to endogenously adopt more structured forms of leadership.

Our field experiments, encompassing more than 5,000 participants, offer valuable insights for researchers as well as practitioners involved in designing incentive schemes for nonroutine analytical team tasks. In particular, we address a prevalent concern among many practitioners of whether monetary incentives impair team performance in tasks that are nonroutine and require thinking outside the box. This concern has been widely propagated in public discourse, notably by best-selling author Daniel Pink through a TED Talk with over 19 million views and his popular book *Drive* (Pink 2009, 2011). Our results alleviate these concerns in the context of teams collaborating on a rich and diverse nonroutine 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 0 (Delfgaauw, Dur, and Souverijn 2020) to 0.90 standard deviations (Hossain and List 2012).[11] Negative effects of incentives have rarely been observed in routine work environments and mostly when pay was low or when performing a routine task could signal prosocial behavior, such as in Hossain and Li (2014), who study the limits of crowding out in a routine data entry task. The authors find that low wages (as compared with no wages) only crowd out participation when a task is framed as a prosocial act but not when it is presented in a work frame or when crowding out does not occur conditional on participation. Complementing

[11] See also table A.1 and the discussion regarding the retail sector and other settings in Delfgaauw, Dur, and Souverijn (2020).

previous work, our findings thus suggest that monetary incentives can provide strong motivations to perform well.

Regarding field experiments involving tasks that are less routine in nature, our work draws parallels to research on incentives for teachers and health practitioners. For both professions, typical tasks require cognitive rather than physical effort and may involve (at least sometimes) novel and unknown problems. As such, we may consider these settings nonroutine and analytical in nature (although it remains unclear whether 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) to 0.31 standard deviations (List, Livingston, and Neckermann 2018; see also table A.1). Evidence regarding incentive pay for health workers is less abundant (Miller and Babiarz 2014), and observed effects sizes are smaller (see app. sec. A.1).

Regarding other nonroutine tasks, our work contributes to the literature on incentives for idea creation (Gibbs, Neckermann, and Siemroth 2017) and creativity (e.g., Ramm, Tjotta, and Torsvik 2013; Gibbs, Neckermann, and Siemroth 2017; Laske and Schroeder 2017; Bradler, Neckermann, and Warnke 2019; Charness and Grieco 2019). These studies 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).[12] One rare exception is a small-scale laboratory experiment by Ramm, Tjotta, and Torsvik (2013), who investigate 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.[13] Our work substantially advances this literature by focusing on a collaboratively 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 by not only increasing workers' effort but also creating a demand for better organizational and leadership structures within teams, which causes additional performance improvements.

The rest of this paper is organized as follows. Section II presents the field setting and the experimental design. Section III provides the main

[12] Laske and Schroeder (2017), Bradler, Neckermann, and Warnke (2019), and Charness and Grieco (2019) study individual production. In Gibbs, Neckermann, and Siemroth (2017), team production is potentially possible, but submitted ideas have fewer than two authors, on average.

[13] Ramm, Tjotta, and Torsvik (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.

results with respect to performance improvements and potential crowding out. Section IV discusses potential mechanisms that shape the efficacy of incentives, and section V provides a more general discussion of our findings. Section VI concludes.

II. Experimental Design and Hypotheses

A. *The Field Setting*

We partner with the company ETR,[14] 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, nonroutine, 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 past few years and can now be found in almost all major cities around the world. 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 room's design 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 exchange and develop innovative and creative ideas to complete the task within a given time limit. If a team manages to complete the task before the allotted time (1 hour) expires, they win. However, if time runs out before the team solves all quests, they lose.

A typical escape room usually features several items, such as desks, shelves, telephones, and books. 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 ones are useful for solving the quests. To illustrate a typical quest in a real-life escape game, we provide a fictitious example.[15] 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 solving

[14] See https://www.exittheroom.de/munich.

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

rebus (word picture) puzzles (see also Kachelmaier, Reichert, and Williamson 2008; Erat and Gneezy 2016).

We conducted our experiments at an ETR facility in Munich. The location offers three rooms with different themes and background stories.[16] 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 they do not manage to do so within 60 minutes, the task is declared incomplete and the activity ends; if they get stuck, they can request hints via radio from the ETR staff. As they can only ask for 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 provide only vague clues regarding the next required step.

ETR provides a rich setting with many aspects of modern nonroutine 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 to solve complex problems. Subjects are required to process stimuli in a way that transcends the usual thinking patterns or are required to use objects in unusual ways. Third, to complete the task, subjects must effectively cooperate as a team. As in other nonroutine team tasks, team members are supposed to provide additional angles to solve 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 nonnegligible 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 when teams are driven by the challenge of solving puzzles and derive enjoyment from making progress in the task, succeeding as fast as possible is clearly desirable.[17] 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

[16] "Zombie Apocalypse" requires teams to find the correct mix of liquids before time runs out (the antizombie potion). In "The Bomb," teams must find a bomb as well as a code to defuse it. 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 tables A.2, A.21).

[17] 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 table A.23).

has a strong incentive to succeed quickly. Finally, the team task is both difficult and nonroutine in nature. This is corroborated by the fact that a substantial fraction of teams fail to finish in 60 minutes (33% of customer teams and 52% of student teams) without incentives, and even a substantial fraction of teams with experienced team members (28% in the field experiment and 50% in the framed field experiment) fail to do so.[18]

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, and analytical and critical thinking) and 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 (Casner-Lotto and Barrington 2006; Jerald 2009; Deming 2017). 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 six.

While these features therefore render escape rooms as an excellent framework for studying the effect of incentives on team performance, the setting is also extremely flexible. Our collaboration with ETR allows us to implement different incentives for more than 700 teams of customers and to also study whether incentives increase performance in a sample of presumably less motivated and exogenously formed teams of student participants (268 teams). The setting's considerable flexibility also enables us to delve into potential mechanisms through which incentives operate (by surveying student teams and conducting an additional natural field experiment that sheds light on the important role of the demand for leadership; see sec. IV).

B. Hypotheses

As customer teams are strongly intrinsically motivated to succeed in the team challenge, there is room for potential motivational crowding out. The theoretical framework outlined in Bénabou and Tirole (2003) formalizes the idea that workers facing incentives may have a distorted perception of their own ability or the task's nature. For example, they may believe that the task is less enjoyable than expected if it needs to be incentivized

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

or that incentives are likely implemented for less intrinsically motivated teams. As such, incentives may increase or decrease performance among intrinsically motivated teams. An increase in performance would mirror the mostly affirmative findings of incentive effects in routine tasks (see, e.g., Erev, Bornstein, and Galili 1993; Lazear 2000; Shearer 2004; Bandiera, Barankay, and Rasul 2005, 2013; Hossain and List 2012; Delfgaauw et al. 2015; Jayaraman, Ray, and de Véricourt 2016; Englmaier, Roider, and Sunde 2017; Friebel et al. 2017), whereas a decrease could substantiate the widely promoted perception that monetary incentives impair team performance when tasks are nonroutine and require thinking outside the box (Pink 2009, 2011). We thus test the following nondirectional hypothesis:

HYPOTHESIS 1. Providing bonus incentives does not affect team performance in the nonroutine task.

Following Bénabou and Tirole (2003), a bonus for extraordinary performance also contains a possible positive signal about a team's ability (because of the ambitious performance goal to which the bonus is tied). Hence, if positive performance effects are observed after the introduction of a bonus, these effects can be caused by the positive team ability signal instead of the reward the bonus provides. Similarly, if crowding out is observed, the actual extent of motivational crowding out due to monetary rewards may be underestimated (because of the compensating effects of the positive signal). The ensuing conjecture is presented in hypothesis 2.

HYPOTHESIS 2. Bonuses with less ambitious performance thresholds lead to more crowding out, while introducing an ambitious reference point (indicating extraordinary performance) without offering a monetary reward improves performance.

The framework by Bénabou and Tirole (2003) also implies that a team's level of intrinsic motivation should mediate incentive effects. For highly intrinsically motivated teams, we expect that apart from causing direct positive incentive effects, extrinsic rewards may reduce intrinsic motivation, whereas for weakly intrinsically motivated teams, such motivational crowding out is less likely. This reasoning implies hypothesis 3.

HYPOTHESIS 3. Teams' intrinsic motivation affects the efficacy of incentives.

In addition to motivational crowding out, incentives may also result in substitutional crowding out (i.e., in a reduction of effort in nonincentivized dimensions; Holmstrom and Milgrom 1991). In particular, bonus incentives for quickly completing a task may alter teams' intrinsic motivation to explore original solutions and instead make them rely more on external help. 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). In our setting, incentives for

speed may reduce teams' effort to explore original solutions (i.e., trying out different approaches on their own and instead asking for hints), particularly when they fail to quickly find the solution themselves.[19] Hypothesis 4 summarizes these arguments.

HYPOTHESIS 4. With bonus incentives, teams are less likely to explore original solutions.

To better understand the roots and causes of our findings, we investigate two particular mechanisms at play. First, independent of crowding out effects on performance, team members' understanding of how effort maps into performance likely affects whether incentives eventually alter outcomes. This seems particularly relevant in nonroutine team work, where subtasks can differ starkly from one another and the inputs by multiple team members aggregate into outputs in a very specific manner. We thus expect the following:

HYPOTHESIS 5. Understanding the production function enhances the performance effects of incentives.

Second, it has been shown that salient incentives may alter team organization (Englmaier, Roider, and Sunde 2017), and in nonroutine tasks, such changes may require efficient leadership. If teams are motivated by the opportunity of receiving an additional bonus payment, incentives may also result in an increased demand for leadership. As leadership has been attributed importance in business, management, economics, and politics (Antonakis et al. 2021), it appears that it is a likely candidate to improve outcomes in nonroutine team tasks. Hence, we hypothesize the following:

HYPOTHESIS 6. Bonus incentives induce demand for leadership, leading to better performance.

C. *Experimental Treatments, Outcome Measures, and Hypotheses Tests*

We conduct the main field experiment with 3,308 customers (722 teams) of ETR Munich and implemented a between-subject design. To test hypothesis 1, our main treatments included 487 teams 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 completed the task in less than 45 minutes.[20] In the Control condition (238 teams), teams were not offered any bonus.

[19] This intuition is also in line with additional survey evidence (see sec. II.E.2) revealing that hints are used to solve difficult puzzles but are perceived as less creative and less original by teams taking few hints.

[20] 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, with about five members, on average). To keep the per-person incentives

We collect 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, the team's gender and age composition,[21] 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.[22] Our primary outcome variable is team performance, which we measure by (1) whether teams complete the task in 45 minutes and (2) the time left upon completing the task. Comparing the Bonus45 with the Control condition allows us to estimate the causal effect of bonus incentives on these objective performance measures.

Notably, the Bonus45 condition includes an ambitious performance threshold (solving the task within 45 minutes rather than in 60 minutes), which may serve as a positive signal for intrinsically motivated workers. To test hypothesis 2, we implement two additional experimental treatments. In 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).[23] That is, teams received the bonus if they completed the task within 60 minutes.[24] In the second additional treatment (Reference Point,

constant in the student sample with three team members (described in sec. II.D.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 app. sec. A.5) 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. To preserve the natural field experiment, we always ensured 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 as either a gain (125 teams) or 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 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 app. sec. A.7.4.

[21] Again, note that to preserve the natural field experiment, we did not interfere with ETR's standard procedures. Thus, we did not explicitly elicit participants' ages. Instead, we estimated each participant's age on the basis of appearance to be (1) below 18 years, (2) between 18 and 25 years, (3) between 26 and 35 years, (4) between 36 and 50 years, or (5) 51 years or older. As requested by the Institutional Review Board, teams with minors were not included in the study.

[22] 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.

[23] Note that in Control, roughly 10% of the teams completed the task within 45 minutes, whereas roughly 67% did so within 60 minutes.

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

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.[25] The performance in Bonus60 as compared with Control allows for 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.

To test hypothesis 3, 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 conceivably less intrinsically motivated. For this purpose, we randomly allocated 804 student participants from the subject pool of the social sciences laboratory at LMU Munich (Munich Experimental Laboratory for Economic and Social Sciences [MELESSA]) into 268 teams. The teams of three students were assigned to treatments Control (88) and Bonus45 (180).[26] Importantly, these participants did not self-select into the escape challenge and were paid to perform the task as part of an economic experiment, which we interpret as implying that they have lower intrinsic motivation.[27] Naturally, both samples differ along a host of dimensions other than intrinsic motivation (e.g., exogenous vs. endogenous team formation, age, or educational background). However, it does not seem obvious to what extent these other differences are likely candidates to explain differential reactions to incentives when testing hypothesis 3.[28]

To test hypothesis 4, we use teams' hint taking as a proxy for whether they explore original solutions. If the bonus (i.e., an incentive for fast completion) reduces teams' effort to try out different approaches, it should become more likely that teams use hints when facing incentives. To test whether knowledge about the production function enhances positive incentive effects (hypothesis 5), we rely on variation in team members' experience with escape challenges.

[25] 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."

[26] Akin to our analyses regarding the natural field experiment, we also pool the two subtreatments Gain45 (90) and Loss45 (90) for the student teams. Appendix sec. A.8 provides additional results on the framing of incentives.

[27] This experiment allowed us to also collect additional data on teams' task perception and team organization (discussed in secs. III and IV).

[28] The intuition that student teams are less intrinsically motivated is also in line with result 4 in sec. III.C.3, which shows that student teams in particular are willing to give up developing original solutions by using more hints when incentivized. Further, other observable characteristics—i.e., dimensions in which student teams may have differed from customer teams, such as cognitive ability (proxied by math and overall grades), relative importance of receiving a monetary reward (proxied by students' income), or task-related abilities (proxied by their field of studies)—do not significantly interact with incentives (see table A.11).

To test hypothesis 6, we use a two-step procedure. First, we compare student teams' demand for leadership between the Bonus45 and the Control condition on the basis of a postexperimental questionnaire. Second, to identify the causal role of an increased demand for leadership, we ran an additional natural field experiment in the exact same setting. In this experiment, we randomly assigned 1,273 regular customers in 281 teams 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. In Leadership, ETR staff highlighted the importance of leadership to succeed in the task and encouraged teams to select a leader from their own group (for the exact wording, see sec. II.D.3).

D. Procedures

1. Natural Field Experiment (Customer Sample)

We conducted the field experiment with ETR customers during regular opening hours from Monday to Friday.[29] We implemented the field experiment's main treatments (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.[30]

Upon arrival, ETR staff welcomed teams of customers as usual, and customers signed ETR's terms and conditions, including its data privacy policy. The staff then explained the rules of the game, and afterward 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.

[29] 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.

[30] All slots in November and December 2015 were fully booked before treatment assignment. According to the provider, fewer than 5% of their bookings are made on the day of an event after the first time slot has ended.

2. Framed Field Experiment (Student Sample)

For the framed field experiment, we invited student participants from MELESSA. Between March and June 2016 and January and May 2017, 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 they 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 occur 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 the second part would include an activity with a participation fee that would be covered by the experimenters.[31]

Upon arriving 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, that is, 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 a future paper.[32] After completing the laboratory part, the experimenters guided the participants to the ETR facility, which is located a 10-minute walk (0.4 mile/650 meters) away from the laboratory. At ETR, each participant was randomly allocated to a team of three members, received the same explanations from ETR staff that were given in the field experiment, and, depending on the treatment, was informed about the possibility of earning a bonus.

For the student sample, we randomized the treatments on the session level (stratifying on rooms), as we made sure that student teams in different sessions on a given day did not encounter each other at the ETR facility. During the performance of the task, the same information about team performance as in the field experiment was collected. Once participants

[31] Appendix sec. A.6 provides a translation of the invitation's text.

[32] These tasks included an elicitation of the willingness to pay for an ETR voucher, an experimental measure of loss aversion (based on Gächter, Johnson, and Herrmann 2022), and a word creation task (developed by Eckartz, Kirchkamp, and Schunk 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, Rentfrow, and Swann 2003), risk preferences (Dohmen et al. 2011), and standard demographics. On average, the subjects spent roughly 30 minutes completing the experimental tasks and questionnaires.

completed the task, they answered questions about the team's behavior and organization, as well as 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 €7.53 on average, with payments ranging from €3.50 to €87.[33]

3. Additional Natural Field Experiment (Leadership)

Between January and March 2018, 1,273 additional regular customers in 281 teams were assigned to one of two experimental conditions: Control-L and Leadership. As before, we randomized on a daily level to avoid treatment spillovers between different teams on site. Participants were not informed that they were taking part in an experiment. The only difference between the conditions was that in Leadership, 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, choose one of you to take the lead and consistently motivate/coordinate the team."[34]

E. Additional Surveys

1. Student Sample

To not interfere with the standard procedures at ETR, we could not run extensive surveys with their customer participants of our natural field experiments. 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): does the task require logical thinking, thinking outside the box, creative thinking, for participants to be concentrated, high effort, and mathematical thinking? Furthermore, we asked whether the task encompassed mostly easy

[33] In one of the laboratory tasks, the student participants further had the chance to win an ETR voucher worth roughly €100. Twenty-six participants actually won 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 2 hours.

[34] The treatment Leadership consisted of two subtreatments that differed only by whether the last sentence stressed the word "motivate" or "coordinate." Since the effects of stressing different leadership functions are not the focus of this paper, see Englmaier et al. (2021) for details.

exercises or to what extent the problems were challenging (both on the same Likert scale).

In addition, we conducted two postexperimental questionnaires 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 19 statements that might capture aspects of team motivation and organization. In questionnaire 2 (which was conducted for a subsample of 375 student participants), we used an additional set of 12 questions based on the concept of team work quality by Hoegl and Gemuenden (2001).[35]

2. Additional ETR Customers

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.[36] Before participating 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 (rank 8) important. Half were asked to rank goals for a hypothetical scenario in which they had the opportunity to win a team bonus of €50 if they completed the task in 45 minutes (bonus condition, $n = 100$). The other half was randomly assigned to a no bonus condition ($n = 101$); that is, they ranked the goals without any bonus being mentioned. After participating in the escape challenge, survey participants had to evaluate by how much they agree with seven statements about hint taking.

3. HR Experts

To estimate the ability of our study to shift priors about the effectiveness of incentives, 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 nonroutine analytical team tasks.[37] Slightly more than half ($n = 203$) were asked about the effectiveness of bonus incentives in escape challenges. We explicitly informed these experts about the nature of the task at hand and asked them to guess how many out of 100 teams (1) would become faster, (2) would become slower, and (3) would do neither once they received the opportunity to earn a bonus. The remaining ($n = 197$) HR experts reported the same numbers for abstract nonroutine

[35] All questions are presented in table 9, where we discuss the results.
[36] Appendix sec. A.15 describes the survey in more detail.
[37] Appendix sec. A.16 describes the survey in more detail.

analytical team tasks (without mentioning escape games). Comparing the assessment of HR experts across tasks allows us to discuss the external validity of our setting.

III. Results

A. *Task Perception and Randomization*

We have previously argued that real-life escape games encompass many features of modern nonroutine 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. Figure 1 shows the mean answers of our postexperimental survey with student participants (see sec. II.E). Participants strongly agree that the task involves logical thinking, thinking outside the box, and creative thinking, in particular as compared with mathematical thinking and easy exercises (signed-rank tests reject that the ratings have the same underlying distribution; all $p < .01$ except for thinking outside the box vs. logical thinking, $p = .16$, and thinking outside the box vs. creative thinking, $p = .02$).

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, on the basis

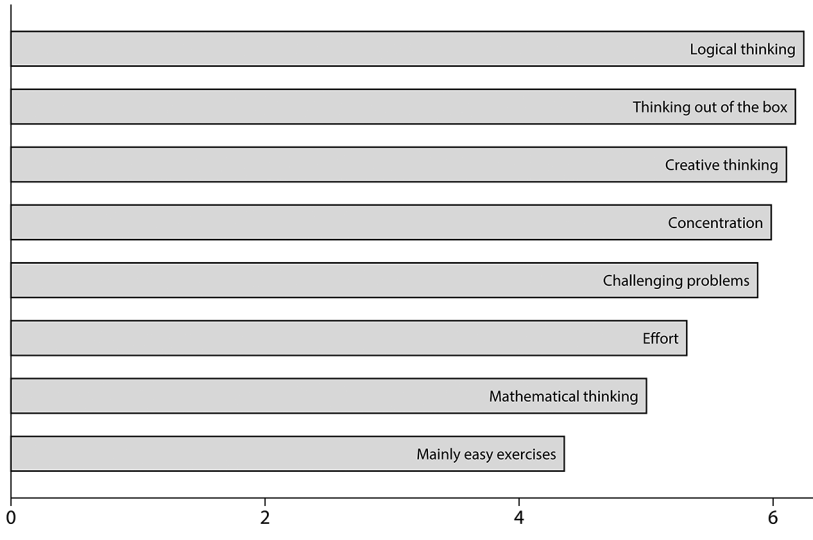

FIG. 1.—Task perception. The figure shows mean answers of $N = 804$ student participants to eight questions concerning the task's attributes. Answers were given on a seven-point Likert scale.

TABLE 1
SAMPLE SIZE AND CHARACTERISTICS

	CONTROL (n = 238)		BONUS45 (n = 249)	
	Mean	Minimum, Maximum	Mean	Minimum, Maximum
Share of men	.52	0, 1	.51	0, 1
	(.29)		(.29)	
Group size	4.53	2, 7	4.71	2, 8
	(1.18)		(1.05)	
Experience	.48	0, 1	.48	0, 1
	(.50)		(.50)	
Private	.69	0, 1	.63	0, 1
	(.46)		(.48)	
English speaking	.12	0, 1	.08	0, 1
	(.32)		(.28)	
Age category ∈ {18–25; 26–35; 36–50; 51+}	{.29; .45; .21; .05}		{.18; .42; .33; .07}***	

NOTE.—All variables except age category represent means on the group level. Experience denotes teams that have at least one member who experienced an escape game before. Private denotes whether a team is composed of private members (1) or whether the team belongs to a team-building event (0). Standard deviations are in parentheses. Age category displays fractions of participants in the respective age category. Asterisks indicate significant differences from Control (using χ^2 tests for frequencies and Mann-Whitney tests for distributions).

*** $p < .01$.

of observables such as the share of men, group size, experience, whether teams were taking part in a private or company event, and whether the team was English speaking. The only characteristic that differs significantly across treatments is the distribution of participants over the age categories guessed by our research assistants (χ^2 test, $p < .01$).[38] 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).

B. Bonus Incentives and Team Performance

We now turn to our primary research question: whether providing bonus incentives improves performance. As previously mentioned, 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 to complete the task.

Figure 2 shows the cumulative distribution of finishing times with and without bonus incentives in the field experiment, with the vertical line marking the time limit for receiving the bonus. The figure indicates that

[38] This does not change when adjusting for multiple hypothesis testing (MHT) according to List, Shaikh, and Xu (2019).

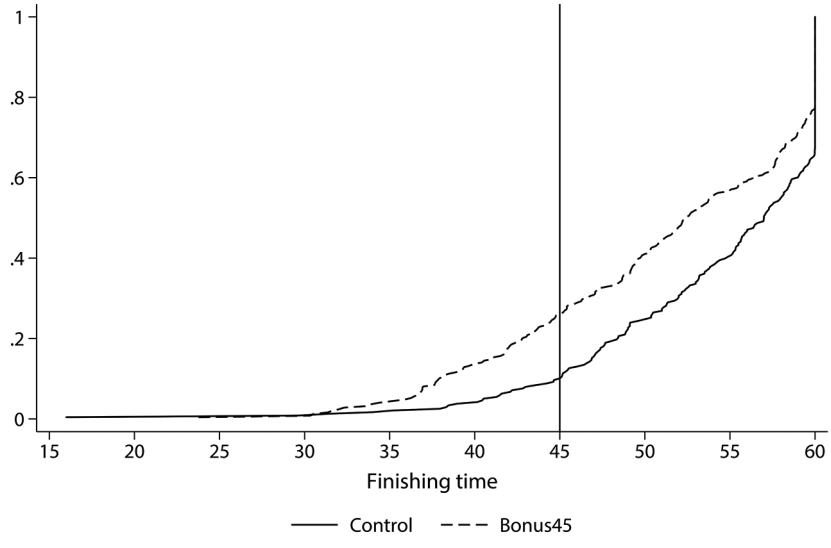

FIG. 2.—Finishing times in Bonus45 and Control in field experiment. The figure shows the cumulative distributions of finishing times with and without bonus incentives. The vertical line marks the time limit for the bonus.

bonus incentives induce teams to complete the task faster. In line with the idea that nonroutine tasks are characterized by a noisy process that translates effort into performance, we observe differences over a large part of the support of the distribution rather than merely around the 45-minute threshold. In Control, only 10% of the teams manage to finish within 45 minutes, whereas in the bonus treatments, more than twice as many teams (26.1%) do so (χ^2 test, $p < .01$). The remaining time upon completion also differs significantly between Bonus45 and Control ($p < .01$, Mann-Whitney test). In Bonus45, teams are about 3 minutes faster than in Control, on average. The positive effect of bonuses on performance is also reflected in the fraction of teams finishing the task within 60 minutes. With bonuses, 77% of the teams finish the task before the 60 minutes expire, whereas in Control, this fraction amounts to only 67% (χ^2 test, $p = .01$). Adjusting p-values for MHT, as suggested in List, Shaikh, and Xu (2019), yields similar results. For further details, see also table A.7 and appendix section A12.1.

In addition to our nonparametric tests, we provide regression analyses that 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

TABLE 2
PROBIT REGRESSIONS: COMPLETED IN LESS THAN 45 MINUTES

	PROBIT (Marginal Effects): COMPLETED IN LESS THAN 45 MINUTES			
	(1)	(2)	(3)	(4)
Bonus45	.165***	.164***	.188***	.151***
	(.024)	(.022)	(.025)	(.041)
Fraction of control teams completing task in less than 45 minutes	.10	.10	.10	.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

NOTE.—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 the base category). Control variables added from col. 2 onward include team size, share of men in a team, a dummy for whether someone in the team has been to an escape game before, dummies for median age category of the team, a dummy for whether the group speaks German, and a dummy for private teams (opposed to company team-building events). Staff fixed effects control for ETR employees present on site and week fixed effects for the week of data collection. All models include the full sample, including weeks that perfectly predict failure to receive the bonus (table A.4 reports regressions from a sample, excluding weeks without variation in the outcome variable). Robust standard errors clustered at the day level are in parentheses.

*** $p < .01$.

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 minutes by 16.5 percentage points. This effect is substantial and equivalent to expanding the team size from four to six members. We add observable team characteristics in column 2,[39] fixed effects for the ETR staff members on duty in column 3, and week fixed effects in column 4. Across all specifications, the coefficients of the bonus treatments are positive and highly significant, indicating that paying bonuses to teams completing a nonroutine task strongly enhances their performance. In table A.5, we also estimate the effects of bonuses on the time remaining upon completing the task, which confirms both the results from the nonparametric tests on the remaining time as well as the results from the probit models in table 2.

Since the incentive rewards completion of the task only within the first 45 minutes, it should become ineffective for the last 15 minutes. In addition,

[39] From the set of characteristics in these and the following analyses, group size, experience with escape games, and the share of men in a team have a positive effect on performance, whereas English-speaking groups perform slightly worse. For more details, see table 8, col. 1.

if incentives crowd out intrinsic motivation to exert effort, we should see a decrease in performance after 45 minutes compared with 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 were true, we should observe the treatment to have a strong effect on the hazard in the first 45 minutes and no or even a negative effect in the last 15 minutes, conditional on covariates.

Table 3 shows the hazard ratios using our usual set of controls and employing cluster robust standard errors. Columns 1–3 estimate the effect on the hazard rate for the first 45 minutes, while columns 4–6 focus on the last 15 minutes. In columns 1 and 4, we present the baseline effect of the treatment without any covariates, which are added in columns 2 and 5. 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 are 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. However, in the last 15 minutes (cols. 4–6), the effect has almost completely vanished. The coefficient on our treatment ranges closely around 1 and is not significantly different from 1 in any specification. Again, the proportional hazard assumption cannot be rejected. Thus, our data reflect two important aspects. First, the treatment indeed increases the likelihood of completing the task in the first 45 minutes but much less

TABLE 3
INFLUENCE OF MAIN BONUS TREATMENT ON HAZARD RATES

	COX PROPORTIONAL HAZARD MODEL: FINISHING TASK					
	First 45 Minutes			Last 15 Minutes		
	(1)	(2)	(3)	(4)	(5)	(6)
Bonus45	2.853***	2.947***	2.914***	1.178	1.251	.841
	(.446)	(.474)	(.844)	(.189)	(.248)	(.180)
p (proportional hazard assumption)	.830	.748	1.000	.800	.686	.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	398	398	398

NOTE.—The table shows hazard ratios from a Cox proportional hazard regression of time elapsed until a team has completed the task on our treatment indicator Bonus45. All models include control variables as well as staff and week fixed effects, as in table 2. Robust standard errors clustered at the day level are in parentheses. 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.

*** $p < .01$.

so in the last 15 minutes. Second, incentives are unlikely to have caused strong feelings of disappointment leading to substantially worse performance after teams failed to achieve the threshold relevant for the bonus payment in our setting. We conclude the following:

RESULT 1. Bonus incentives increase team performance in the nonroutine task.

C. *Potential Crowding Out of Intrinsic Motivation*

Importantly, the results from our field experiment demonstrate that bonus incentives substantially improve team performance among teams with strong intrinsic motivation. As such, the monetary reward of the bonus appears to outweigh potential negative effects due to the crowding out of intrinsic motivation. However, in Bonus45, the bonus incentive was tied to an ambitious performance threshold (45 minutes) that only 10% of teams in Control could achieve. Hence, it is crucial to investigate whether bonuses also work when they are not coupled with ambitious performance thresholds (see hypothesis 2).

Furthermore, we aim to explore the robustness of incentive effects among a sample of less intrinsically motivated teams. Doing so allows us to go beyond merely analyzing the potential net effect of incentives and potential crowding out. In particular, observing similar effect sizes among differently intrinsically motivated teams would likely suggest that the net effect aligns with the pure positive effect of bonus incentives (see hypothesis 3). Finally, we seek to uncover whether crowding out can be observed in the form of substitution of (multidimensional) effort by shedding light on teams' exploration behavior (i.e., hint taking; see hypothesis 4).

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 refer to figure 3. This figure displays the cumulative distribution of finishing times in conditions Control, Reference Point, Bonus60, and Bonus45. It suggests that monetary rewards reduce the amount of time teams need to finish the task, even when coupled with a less ambitious performance goal of 60 minutes (Bonus60 vs. Control, Mann-Whitney test, $p = .05$; Bonus45 vs. Control, Mann-Whitney test, $p < .01$; Bonus45 vs. Bonus60, Mann-Whitney test, $p = .24$). Further, we do not observe that the ambitious reference point independently improves performance, as the cumulative distribution of remaining times

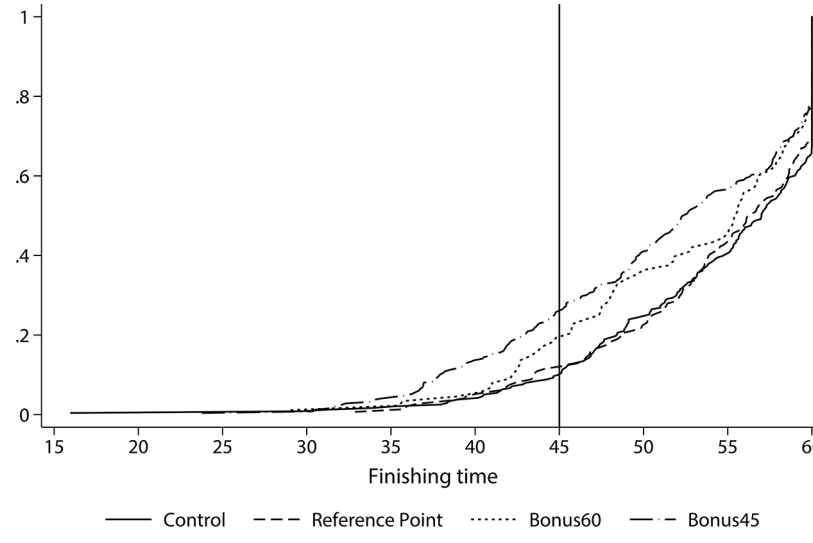

FIG. 3.—Finishing times for all treatments in field experiment. The figure shows the cumulative distribution of finishing times of Bonus45 (pooled), Bonus60 (pooled), Reference Point, and Control. The vertical line marks the time limit for the bonus in the Bonus45 condition.

in Reference Point almost perfectly overlaps with the cumulative distribution function in Control (Mann-Whitney test, $p = .78$).[40]

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, column 2 adds our set of control variables, column 3 adds staff fixed effects, and column 4 adds week fixed effects. The regressions show that monetary incentives significantly increase the probability of finishing within 45 minutes, whereas the reference treatment does not.[41] It also becomes apparent that this finding is robust to adding covariates and fixed effects.

Moreover, a postestimation Wald test rejects the equality of coefficients of Bonus60 and Reference Point in all specifications (cols. 1–4, $p < .1$). Similarly, the coefficient of Bonus45 is significantly larger than the coefficient of Reference Point in all specifications ($p = .07$ in col. 4, $p < .01$ in all other specifications). Equality of coefficients of Bonus60

[40] The results point in a similar direction when adjusting for MHT following the approach suggested in List, Shaikh, and Xu (2019; see app. sec. A.12.1 for details).

[41] Table A.6 confirms these findings for remaining time as the dependent variable.

TABLE 4
PROBIT REGRESSIONS: COMPLETED IN LESS THAN 45 MINUTES (All Treatments)

	PROBIT (Marginal Effects): COMPLETED IN LESS THAN 45 MINUTES			
	(1)	(2)	(3)	(4)
Bonus45	.160***	.157***	.164***	.108***
	(.023)	(.022)	(.026)	(.035)
Bonus60	.105**	.102***	.105***	.127**
	(.041)	(.038)	(.039)	(.051)
Reference Point	.025	.023	.011	.020
	(.032)	(.035)	(.039)	(.039)
Bonus45 = Bonus60	[.151]	[.095]	[.120]	[.752]
Bonus45 = Reference Point	[.000]	[.000]	[.000]	[.073]
Bonus60 = Reference Point	[.066]	[.059]	[.033]	[.024]
Fraction of control teams completing task in less than 45 minutes	.10	.10	.10	.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

NOTE.—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. All models include control variables as well as staff and week fixed effects, as in table 2. Robust standard errors clustered at the day level are in parentheses. Square brackets indicate p-value of Wald test for equality of coefficients.

** $p < .05$.

*** $p < .01$.

and Bonus45 can be rejected for only one of the specifications (col. 2, $p = .095$). 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.

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 turn to our 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 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.[42]

Across both treatments, student teams do not differ significantly in any observed characteristic. The average share of men in Bonus45 (0.43) is not significantly different from Control (0.45; Mann-Whitney test, $p = .31$), and neither is the share of teams with at least one experienced member (0.36 vs. 0.36, χ^2 test, $p = .90$) or teams' average age (22.96 vs. 23.18 years, Mann-Whitney test, $p = .72$). Nevertheless, we control for team characteristics in our regression analyses.

Analogous to the analysis in the customer sample, we study the treatment effects on team performance by analyzing the fraction of the teams completing the task within 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, serving as the student sample counterpart to figure 2. While student teams perform, on average, substantially worse than the ETR customer teams, the bonus incentives prove to be similarly effective for the student teams.[43]

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% of teams manage to complete the task within 45 minutes, whereas only 5% do so in Control (χ^2 test, $p = .08$). The fraction of teams finishing the task within 60 minutes is also significantly larger under bonus incentives. With bonuses, 60% of the teams finish the task before the 60 minutes expire, whereas in Control, this fraction amounts to 48% (χ^2 test, $p = .06$). Further, with bonus incentives, teams are, on average, about 3 minutes faster than in Control, and Mann-Whitney tests reject

[42] As discussed in sec. II.C, ETR customer teams 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 are, on average, younger (23.03 years), slightly less likely to be male (44%), and less experienced in escape games (36% of the student teams had at least one member with escape game experience). As shown in table A.11, these characteristics (apart from experience) do not significantly relate to teams' probability of receiving the bonus.

[43] Given the differences in completion rates at 45 minutes in the Control condition across student and customer teams, we provide further analyses assessing the treatment effects by the minute using a Cox proportional hazard model, which additionally controls for team characteristics, staff, and week fixed effects. Figures A.1 and A.2 reveal that students' conditional likelihood of success remains low until the 50-minute mark in Control and then sharply increases. In contrast, for customer teams in Control, we find a gradual increase from minute 35 onward, indicating a richer heterogeneity among customer teams' performance. With incentives (Bonus45), the hazard rates among both student and customer teams steadily increase from the 35-minute mark onward. This is in line with the idea that teams provide more effort early on (in the hope of receiving the bonus payment) and do not completely slack after the 45-minute mark has passed (see also the analyses in table 3). Hence, incentives increase the likelihood of finishing early in both samples, and their efficacy does not seem to strongly depend on the underlying heterogeneity in teams' performances without incentives.

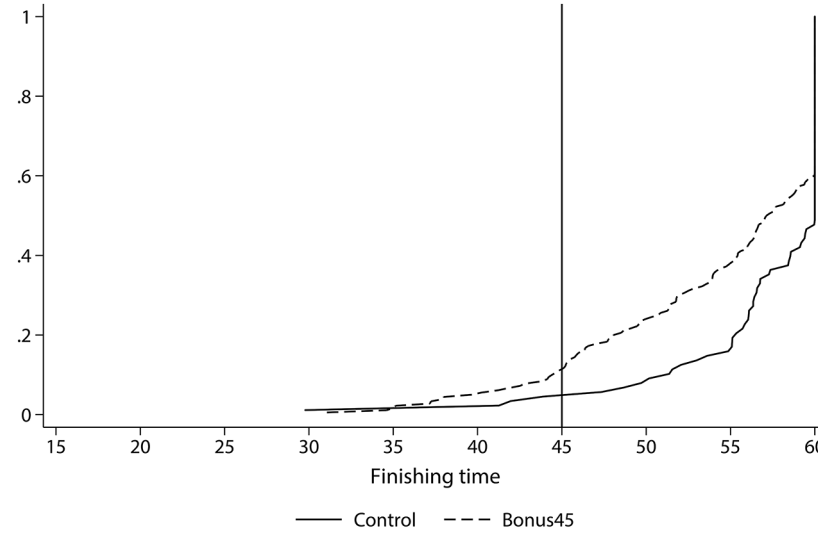

FIG. 4.—Finishing times in Bonus45 and Control in framed field experiment (student sample). The figure shows the cumulative distributions of finishing times with and without bonus incentives in the framed field experiment. The vertical line at 45 minutes marks the time limit for the bonus.

that finishing times in the control condition come from the same underlying distribution as finishing times under bonus incentives (Mann-Whitney test, $p < .01$).[44] These results are also robust to adjusting p-values for MHT, as suggested in List, Shaikh, and Xu (2019; see app. sec. A.12.2 for more details).

In addition to the nonparametric tests, we run regressions analogously to the analysis for the customer sample. As before, we control for the share of men in a team, average age, and experience with escape games.[45] Columns 1–4 of table 5 report the results from probit regressions on the probability of completing the task within 45 minutes. Column 1 uses only the treatment dummy and shows that bonus incentives significantly increase the probability of completing the task within 45 minutes. The positive effect of the bonus incentives is robust to controlling for background characteristics (col. 2), staff fixed effects (col. 3), and week fixed effects (col. 4). Overall, the probit regression results reinforce our nonparametric findings: offering bonuses increases team performance.

[44] Table A.8 summarizes these findings and provides further details with respect to the framing of incentives.

[45] 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 (Marginal Effects): COMPLETED IN LESS THAN 45 MINUTES				Pooled OLS
	(1)	(2)	(3)	(4)	(5)
Bonus45	.075*	.073*	.075*	.079**	.086***
	(.042)	(.041)	(.039)	(.037)	(.030)
Field					.290*
					(.151)
Bonus45 × field					.083
					(.059)
Fraction of control (student) teams completing task in less than 45 minutes	.05	.05	.05	.05	.05
Control variables	No	Yes	Yes	Yes	Yes
Staff fixed effects	No	No	Yes	Yes	Yes
Week fixed effects	No	No	No	Yes	Yes
Observations	268	268	268	268	755

NOTE.—Columns 1–4 show average marginal effects from probit regressions of whether a team completed the task within 45 minutes on our treatment indicators (with Control as the base category). Column 5 reports coefficients from a linear regression including both the student and the customer sample. Control variables added from col. 2 onward include the share of men in a team, a dummy for whether someone in the team has been to an escape game before, and average age of the team. Staff fixed effects control for ETR employees present on site, and week fixed effects control for the week of data collection. All models include the full sample, including weeks that perfectly predict failure to receive the bonus (table A.10 reports regressions from the student sample, excluding weeks without variation in the outcome variable). Robust standard errors clustered at the session level are in parentheses.

* $p < .10$.
** $p < .05$.
*** $p < .01$.

Column 5 reports results from a linear regression, in which we pool both samples and test for the interaction of incentives and the specific sample. The results show no differential effect of incentives for the customer versus the student sample. Furthermore, for the student sample, the positive effect of bonus incentives is reflected qualitatively in the analyses of the time remaining (see table A.9). These results emphasize that a crowding out of intrinsic motivation does not seem to strongly distort the pure effect of incentives.[46] We summarize these findings as follows:

[46] As previously discussed, we do not find it obvious to what extent any sample differences in characteristics other than intrinsic motivation would affect performance. Given that we do not observe differences in treatment effects across the samples, any differences in other (un)observable characteristics between the groups could influence the result only if they exactly canceled out the effects introduced by differences in intrinsic motivation, which appears unlikely. Additionally, as table A.11 shows, no other observed characteristics interact with the performance effect among the student participants.

RESULT 3. Incentives are similarly effective among teams that self-selected into the task (customer teams) and teams assigned to the task by us (student teams).

3. Bonus Incentives and Team Willingness to Explore

To test hypothesis 4, we next analyze how many out of the five possible hints teams request under the different treatment conditions as well as whether they are more likely to take hints earlier in the presence of incentives.[47]

Table 6 shows the number of hints taken across samples and treatments. For teams that 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 with Control, suggesting that incentives reduce these student teams' willingness to explore original solutions.[48]

To capture potential heterogeneity across teams, we report the fractions of teams requesting zero, one, two, three, four, or five hints for the customer sample in figure 5*A* and for the student sample in figure 5*B*. 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, $p = .11$). Figure 5*B* depicts the same histogram for the framed field experiment with student participants. It becomes apparent that teams that did not self-select into the task are much more likely to take hints when facing incentives (χ^2 test, $p = .029$). Roughly 75% of these teams take four or five hints when facing incentives as compared with 59% doing so in Control. Ordinary least squares (OLS) regression analyses for hint taking including additional controls (see table 7, cols. 1 and 3) confirm these results.[49]

[47] Appendix sec. A.11 provides 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.

[48] Note that a similar picture arises if we standardize the task's length to account for different completion times by customer and student teams. We convert the time the hint was taken as a fraction of the total game time (either actual time of completion or 60 minutes, in case teams did not complete the task). Figures A.3 and A.4 plot the average fraction of hints taken conditional on the share of time elapsed in the customer and student sample across treatments. The figures show that incentives leave hint taking among customer teams virtually unchanged, whereas student teams seem to use more hints when facing incentives after around 20% of the standardized length of the game has passed.

[49] An ordered probit regression yields qualitatively similar results; see table A.13.

TABLE 6
HINTS REQUESTED IN FIELD EXPERIMENT AND FRAMED FIELD EXPERIMENT

Experiment	Control	Bonus45
Within 60 minutes:		
Field (487 groups)	2.92	3.10
	(1.55)	(1.34)
Framed field (268 groups)	3.74	4.11***
	(1.04)	(.98)
Within 45 minutes:		
Field (487 groups)	1.97	2.36***
	(1.22)	(1.15)
Framed field (268 groups)	2.33	3.17***
	(.93)	(1.04)

NOTE.—The table summarizes the mean number of hints taken across treatments in the field experiment and the framed field experiment (standard deviations in parentheses). Asterisks indicate significant differences from Control (using Mann-Whitney tests). Teams in the framed field experiment take more hints within 60 minutes (Control: $p < .01$; Bonus45: $p < .01$) and within 45 minutes (Control: $p = .013$; Bonus45: $p < .01$). p-values of nonparametric comparisons between Gain45 and Loss45 are larger than 0.10 for both experiments.

*** $p < .01$.

When we focus only on hints taken within the first 45 minutes, nonparametric tests indicate significant differences across treatments for both samples, but again, the effect is much stronger for student teams that we assigned to the nonroutine task (customers: χ^2 test, $p < .01$; students: χ^2 test, $p < .01$). Regression analyses using additional controls and fixed effects imply that these teams take, on average, 0.808 more hints within the first 45 minutes when facing incentives, whereas customer teams take, on average, only 0.186 more hints (cols. 2 and 4 of table 7). Hence, the nonparametric results for the student sample remains largely unchanged,

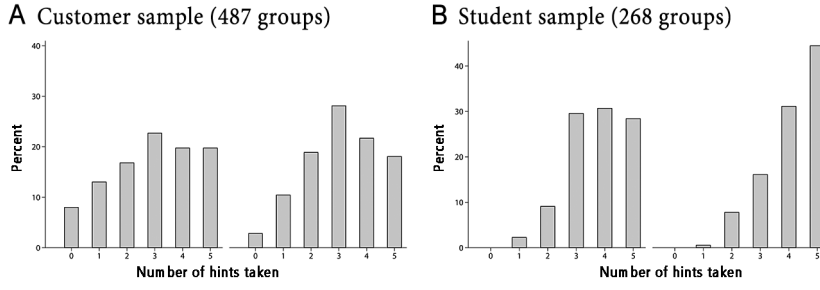

FIG. 5.—Hints requested across samples and treatments. The figure shows histograms of hints taken across samples. *A* depicts the fractions of customer teams choosing 0, 1, 2, 3, 4, or 5 hints in Control (*left*) and Bonus45 (*right*). *B* shows the fractions of student teams.

TABLE 7
OLS Regressions: Number of Hints Requested

	OLS: Number of Hints Requested					
	Field Experiment		Framed Field Experiment		Pooled	
	Within 60 Minutes (1)	Within 45 Minutes (2)	Within 60 Minutes (3)	Within 45 Minutes (4)	Within 60 Minutes (5)	Within 45 Minutes (6)
Bonus45	.098	.186	.343**	.808***	.357***	.829***
	(.183)	(.134)	(.136)	(.122)	(.117)	(.119)
Field					−2.589***	−1.917***
					(.603)	(.385)
Bonus45 × field					−.297	−.674***
					(.217)	(.182)
Constant	4.037***	1.770***	5.391***	4.236***	4.994***	3.363***
	(.442)	(.469)	(.650)	(.698)	(.439)	(.416)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Staff fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Week fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	487	487	268	268	755	755

Note.—The table shows coefficients from OLS regressions of the number of hints requested within 60 or 45 minutes regressed on our treatment indicator Bonus45 (pooled). The sample is restricted to the (natural) field experiment in cols. 1 and 2 and the framed field experiment in cols. 3 and 4. Columns 5 and 6 include both samples. Field is a dummy equal to 1 for the (natural) field experiment. Controls and fixed effects are identical to previous tables. Robust standard errors clustered at the day (for the field experiment) or session (for the framed field experiment) level are in parentheses.

** $p < .05$.

*** $p < .01$.

whereas the positive effect observed in our nonparametric analyses becomes small and statistically insignificant for the customer sample.

In columns 5 and 6, we pool the data from the two samples and study whether there is a significant difference in the reaction to the bonus incentive (in terms of hint taking) in the customer as compared with the student sample. While students in the incentive condition do not generally react substantially differently to the incentive by taking more hints (see col. 5), bonus incentives indeed substantially increase their hint-taking behavior as long as the bonus threshold can still be achieved (i.e., within the first 45 minutes; see col. 6).

Overall, our results align with the conclusion that intrinsic motivation and incentives interact complexly when teams can choose whether to explore original and innovative solutions on their own.[50] Incentives increase the hint-taking behavior of teams that did not self-select into the task, indicating a substitution of effort due to incentives, in line with the

[50] These findings complement recent work on incentive effects in meaningful routine tasks (Kosfeld, Neckermann, and Yang 2017).

multitasking framework by Holmstrom and Milgrom (1991). However, such substitution is much less prevalent among intrinsically motivated customer teams, aligning with the idea that these teams may derive utility from progressing on their own and hence take fewer hints.

To understand whether this idea is reflected in teams' perceptions, we turn to our additional survey among ETR customers and analyze how teams' perceptions differ conditional on their own hint-taking behavior. While both teams that take few (less than three) or many (three or more) hints similarly agree that hints are used to solve difficult puzzles (χ^2 test, $p = .71$), we observe that teams taking few hints perceive hint taking more negatively, particularly as less creative (χ^2 test, $p < .01$), less original (χ^2 test, $p < .01$), and less fun (χ^2 test, $p < .01$).[51]

An alternative explanation for reduced substitution among intrinsically motivated teams (as compared with 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 present. However, this should likely result in less hint taking among teams in the bonus condition as compared with 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 with less ambitious) performance thresholds (3.09 hints in Bonus45 vs. 3.26 hints in Bonus60, χ^2 test, $p = .84$), rendering compensating behavior unlikely.

We summarize our findings in result 4.

Result 4. As long as the bonus can still be achieved (i.e., within the first 45 minutes), incentives increase hint taking by teams hired to perform the task (student teams). This effect is much smaller and statistically insignificant among teams that chose to perform the task (customer teams).

IV. Mechanisms

Our results have shown that incentives causally and unambiguously improve team performance but have not yet established how they improve performance. We aim to provide insights on likely mechanisms through two distinct avenues. First, to better understand what distinguishes teams that do respond to incentives from those that do not, we discuss whether any particular observable team features interact with the observed efficacy of incentives. Second, we investigate how incentives affect behavior, particularly team organization. Postexperimental survey responses identify

[51] For further details on the survey, see app. sec. A.15.

increased demand of leadership as a potential channel, which we subsequently investigate using our additional natural field experiment.[52]

A. *When Do Incentives Work?*

We first investigate whether the efficacy of incentives for solving the task within 45 minutes interacts with customer teams' observable characteristics in table 8.[53] The results do not contain significant interactions with the teams' gender share (col. 2), team size (col. 3), teams' language (col. 6), or whether teams participated as part of a company event (col. 5). This suggests 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 (col. 4). Experienced members possess more knowledge about how team effort translates into team success, which could enhance the effects of incentives. We find a positive, economically and statistically significant interaction of bonus incentives and experience. Our estimates imply that the positive bonus effect is about 1.5 times larger for experienced teams. 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 reflected in teams' remaining times, where the bonus tends to be more effective for experienced teams, though not at conventional significance levels ($p = .10$; see col. 4 in table A.3). For remaining times, we also find that a higher share of men relates positively to performance but decreases the effectiveness of incentives (possibly because of 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 fig. A.5). This result aligns 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 probability 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. The effect of bonus incentives is larger for teams with experienced team members.

[52] Additionally, in app. sec. A.14, we provide a broader discussion of the dimensions along which incentives may change behavior within teams, including even more additional surveys and an additional laboratory experiment.

[53] Table A.3 provides results for teams' remaining times.

TABLE 8
LINEAR PROBABILITY MODEL: COMPLETED IN LESS THAN 45 MINUTES

	OLS: COMPLETED IN LESS THAN 45 MINUTES					
	(1)	(2)	(3)	(4)	(5)	(6)
Bonus45	.172***	.200***	.023	.120**	.130**	.169***
	(.050)	(.071)	(.122)	(.057)	(.056)	(.047)
Share of men	.102*	.130***	.102*	.100*	.105*	.103*
	(.055)	(.048)	(.055)	(.054)	(.056)	(.058)
Group size	.056***	.056***	.042**	.057***	.055***	.056***
	(.017)	(.017)	(.017)	(.017)	(.017)	(.017)
Experience	.125***	.126***	.126***	.058*	.124***	.125***
	(.031)	(.031)	(.032)	(.032)	(.031)	(.031)
Private	.040	.039	.039	.036	−.001	.039
	(.041)	(.042)	(.042)	(.041)	(.049)	(.041)
English speaking	−.115*	−.117*	−.113*	−.114*	−.117*	−.129***
	(.060)	(.062)	(.062)	(.060)	(.059)	(.044)
Bonus45 × share of men		−.055				
		(.128)				
Bonus45 × group size			.031			
			(.025)			
Bonus45 × experience				.132**		
				(.051)		
Bonus45 × private					.077	
					(.056)	
Bonus45 × English speaking						.027
						(.139)
Constant	−.177	−.192	−.109	−.179	−.163	−.172
	(.132)	(.133)	(.142)	(.132)	(.133)	(.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

NOTE.—The table shows coefficients from a linear probability model. The dependent variable is a 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 are in parentheses.

* $p < .10$.
** $p < .05$.
*** $p < .01$.

B. Performance and Team Organization

Table 9 shows the results from questionnaires 1 and 2, reporting uncorrected *p*-values as well as MHT-adjusted *p*-values with 31 outcomes, following List, Shaikh, and Xu (2019). Panel A shows that overall, incentives 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, $p < .01$).[54] At the same time, similar to teams in Control, treated teams strongly agree with the statement "I would like to perform a similar task again" (Mann-Whitney test,

[54] 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 9
ANSWERS TO POSTEXPERIMENT QUESTIONNAIRES

	Control	Bonus45	p	MHT-Adjusted p
	A. Questionnaire 1 (n = 804)			
1. The team was very stressed.	3.57	4.13***,†††	.000	.000
2. One person was dominant in leading the team.	2.60	2.86**	.028	.396
3. We wrote down all numbers we found.	5.64	5.50**	.044	.991
4. I was dominant in leading the team.	2.64	2.87**	.053	.520
5. We first searched for clues before combining them.	4.58	4.39	.107	.899
6. We exchanged many ideas within the team.	5.87	5.74	.119	.904
7. When we got stuck, we let as many team members try as possible.	5.43	5.28	.143	.914
8. The team was very motivated.	6.14	6.27	.221	.881
9. We communicated a lot.	5.78	5.88	.227	.982
10. All team members exerted effort.	6.24	6.37	.242	.850
11. Our notes were helpful for finding the solution.	5.50	5.43	.413	.999
12. I was able to present all my ideas to the group.	5.95	5.93	.406	.991
13. We were well coordinated within the group.	5.73	5.80	.606	.997
14. I was too focused on my own part.	2.88	2.83	.763	1
15. We made our decisions collectively.	5.51	5.58	.867	.999
16. I would like to perform a similar task again.	6.30	6.28	.876	.985
17. Our individual skill sets complemented each other well.	5.65	5.68	.891	.998
18. We had a good atmosphere in the team.	6.30	6.37	.929	.992
19. All team members contributed equally.	5.97	6.00	.956	.999
	B. Questionnaire 2 (n = 375)			
1. To what extent did you want someone to take the lead?	2.67	3.32***,†††	.000	.009
2. How well was the team led?	3.85	4.21**	.036	.400
3. How deeply did you think about the problems?	6.00	5.79	.111	.553
4. To what extent did you follow ideas that were not promising?	5.02	4.79	.173	.772
5. To what extent did you develop a team spirit?	5.54	5.80	.168	.760
6. How well were individual tasks and joint strategy coordinated?	3.28	3.51	.183	.914
7. How well did you leverage team members' individual potential?	5.14	4.94	.217	.890
8. How much did you help each other when someone was stuck?	5.70	5.58	.217	.994
9. How intensely did you search the room for clues?	6.31	6.22	.515	.994
10. How much effort did all the team members exert?	5.98	5.96	.600	.908
11. How much did you communicate about procedures?	5.30	5.35	.883	1
12. How willing were team members to accept the help of others?	5.80	5.85	.892	1

NOTE.—The table reports answers to our postexperiment 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 with the statement (1) to completely agreeing (7) in questionnaire 1 and from very little (1) to very much (7) in questionnaire 2. Asterisks indicate significant differences from Control using Mann-Whitney tests, and daggers indicate significant differences when adjusting for MHT (concerning 31 outcomes), according to List, Shaikh, and Xu (2019).

** $p < .05$.

*** $p < .01$.

††† $p < .01$.

$p = .88$, MHT-adjusted $p = .99$), suggesting that incentives cause positive rather than negative stress among the team members. Second, participants in the incentive treatment tend to agree more with the statement that "one person was dominant in leading the team" (Mann-Whitney test, $p = .03$, MHT-adjusted $p = .40$) as well as with the statement "I was dominant in leading the team" (Mann-Whitney test, $p = .05$, MHT-adjusted $p = .52$). However, both of these statements lack statistical significance when adjusted for MHT.

The results from questionnaire 2 in panel B of table 9 mirror the answers from questionnaire 1. Teams facing incentives wish for more leadership (Mann-Whitney test, $p < .01$) and tend to report that teams were better led (Mann-Whitney test, $p = .04$, MHT-adjusted $p = .40$). However, the latter fails to reach conventional significance levels when adjusting for MHT. Overall, both questionnaires suggest that incentives may change the way teams are organized, indicating that incentives may lead to an endogenous emergence of (a demand for) team leaders. This inference is also supported by an alternative approach to adjust for MHT, where principal component factor analyses is used for dimensionality reduction, following the Kaiser-Guttman rule (see Loehlin and Beaujean 2016). We apply this method separately for questionnaires 1 and 2 in 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).[55] We find that general team collaboration does not significantly differ across treatments (Mann-Whitney test: $p = .76$) and neither does dominance (Mann-Whitney test: $p = .11$). However, incentives tend to increase team cohesion (Mann-Whitney test: $p = .07$) and intensity (Mann-Whitney test: $p < .01$) but decrease documentation (Mann-Whitney test: $p = .02$).

Regarding questionnaire 2, we retain three factors that we term as cooperative (factor 1), leadership (factor 2), and struggling (factor 3).[56] Cooperative behavior (factor 1) does not significantly differ across treatment conditions (Mann-Whitney test: $p = .34$). Leadership (factor 2) is significantly more pronounced with incentives (Mann-Whitney test: $p < .01$). Struggling in teams (factor 3) tends to be lower with incentives but statistically insignificantly so (Mann-Whitney test: $p = .26$). Overall, both analyses indicate that incentives appear to change team organization and stimulate the demand for—and the emergence of—leadership.

[55] Items from questionnaire 1 that load heavy on factor 1 are 5, 6, 7, 9, 13, 15, and 18. Items loading heavy on factor 2 are 8, 10, 12, 16, 17, and 19. Items loading heavy on factor 3 are 2 and 4, those loading heavy on factor 4 are 3 and 11, and those loading heavy on factor 5 are 1 and 14.

[56] 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, and items that load high on factor 3 are 3, 4, and 9.

C. *The Causal Effect of Leadership*

To investigate the causal demand of an increased demand for leadership, we ran an additional natural field experiment in which teams were either randomly encouraged to choose a leader (Leadership) or not (Control-L; see also sec. II.D.3). Figure 6 shows the cumulative distribution functions of finishing times across both conditions. Teams in the Leadership treatment condition clearly perform better than those in the Control-L condition. Specifically, in Leadership, 63% of teams finish the task within the time limit of 60 minutes, whereas only around 44% do so in Control-L (Pearson χ^2 test: $p < .01$). 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: 3 minutes and 10 seconds in Control-L and 5 minutes and 29 seconds in Leadership; Mann-Whitney test: $p < .01$).

These nonparametric results are confirmed by a series of probit regressions, where we incrementally introduce additional control variables and fixed effects as in table 2. In table 10, we estimate the average marginal effect of Leadership on the probability of completing the task within 60 minutes. As before, we cluster standard errors at the daily level, which also corresponds to 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

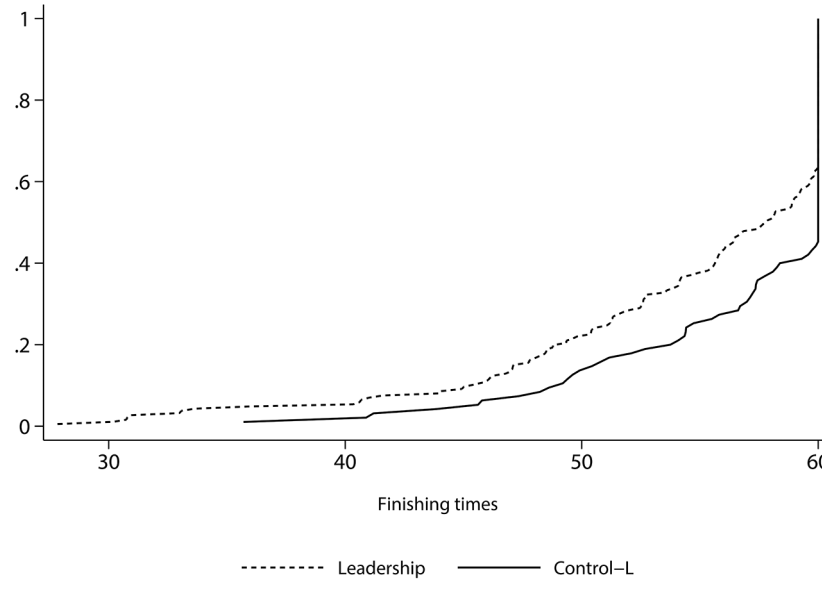

FIG. 6.—Leadership: cumulative distribution functions of finishing time. The figure shows the cumulative distribution of finishing times for teams in Leadership and Control-L.

TABLE 10
PROBIT REGRESSIONS: LEADERSHIP, COMPLETED IN LESS THAN 60 MINUTES

	COMPLETED WITHIN 60 MINUTES			
	(1)	(2)	(3)	(4)
Leadership	.182***	.187***	.185***	.168***
	(.051)	(.052)	(.065)	(.051)
Fraction of teams in Control-L completing task in 60 minutes	.442	.442	.442	.442
Controls	No	Yes	Yes	Yes
Staff fixed effects	No	No	Yes	Yes
Week fixed effects	No	No	No	Yes
Observations	281	281	281	281

NOTE.—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 the base category). Each column indicates whether team controls (group size, share of men, experience with escape games, median age, language spoken, private vs. team-building events, actively taken walkie-talkie) as well as staff and week fixed effects are included. Standard errors clustered at the daily level are in parentheses.

*** $p < .01$.

estimated average marginal effect amounts to an increase of 17 percentage points as compared with Control-L, implying a relative increase in the fraction of successful teams by about 38%.

In 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 increases the remaining time upon task completion by, on average, 2 minutes and 48 seconds.[57] 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. Exogenously shifting the demand for leadership results in substantial performance improvements.

[57] Note that the magnitudes are hardly comparable with the results presented in tables 2 and A.5, as incentives targeted task completion after 45 minutes, whereas the leadership intervention targeted completion only at the 60-minute mark. The cleanest comparison for the case of incentives would be to regress the remaining times or the likelihood of completion in 60 minutes on the Bonus60 treatment. Doing so in the full specification results in a marginal effect of an additional 2 minutes and 44 seconds of remaining time and a 12.5 percentage point increase in completion probability. The latter effect is somewhat lower, albeit not significantly so, than the effect of leadership; however, teams in the different control groups exhibited varying levels of success (0.442 in Control-L vs. 0.67 in Control). This suggests that leadership possibly had a larger scope to improve performance on the extensive margin. Therefore, the emergence of leadership seems to have a comparable potential for improving performance to that of offering bonus incentives.

V. Discussion

Our results demonstrate that bonus incentives have sizable positive effects on team performance in both the natural and the framed field experiments. Building upon important work by Maniadis, Tufano, and List (2014), we investigate how much our findings should update our beliefs that incentives truly increase performance in our task. To do so, we calculate poststudy probabilities (PSPs) conditional on different priors. $\text{PSP} = (1 - \beta)\pi/[(1 - \beta) + \alpha(1 - \pi)]$, where π denotes the probability of a given prior and $(1 - \beta)$ denotes 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 11, where the rows display increasing priors and the columns reflect different levels of power. Column 1 shows posteriors given a 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 upward. Starting from priors as low as $\pi = 0.10$, which indicate a strong disbelief in any effect, the posteriors

TABLE 11
POSTSTUDY PROBABILITIES

	χ^2 Tests on Success Dummy (45 and 60 Minutes) in Framed Field (1)	χ^2 Tests on Success Dummy (45 Minutes) in Natural Field (2)	χ^2 Tests on Success Dummy (60 Minutes, Natural Field) and t-Tests on Remaining Time (Natural and Framed Field) (3)
Achieved power	.45	.70	.95
	Posterior		
Prior probability:			
.05	.32	.42	.50
.1	.50	.61	.68
.2	.69	.78	.83
.4	.86	.90	.93
.6	.93	.95	.97
.8	.97	.98	.99
.9	.99	.99	.99

NOTE.—The table reports PSPs (Maniadis, Tufano, and List 2014) for different combinations of prior probabilities and achieved power. The levels of power in cols. 1–3 correspond to the achieved power in terms of statistical tests (t-tests and χ^2 tests) for our primary outcomes. We achieve a power of about 0.95 for t-tests on the remaining time in the natural and framed field experiment as well 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 within 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 within 45 or 60 minutes amounts to 0.45.

reflect equal probabilities of both outcomes (PSP = 0.50). Priors of $\pi > 0.10$ yield posteriors strongly in favor of our result.

Column 2 shows posteriors for a power of $(1 - \beta) = 0.7$, which corresponds to our binary outcome variable on succeeding within 45 minutes for the natural field experiment. Column 3 reports posteriors for a power of $(1 - \beta) = 0.95$, which we achieve for our binary outcome variable on succeeding in 60 minutes in the natural field experiment, as well as for *t*-tests on the remaining time in both the framed and the natural field experiment. Both columns 2 and 3 show that even moderate to high disbelief converts into posteriors strongly favoring an effect to exist.

To establish a realistic prior, we turn to our survey with HR experts. On average, these experts believed that 40.38% of teams would improve in performance, 23.33% of teams would decline, and outcomes for 36.29% of teams would remain unchanged. As table 11 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.[58] These calculations emphasize the strong updating that decision makers should undergo as they learn about the results from our study.[59]

Our series of large-scale field experiments constitutes, to the best of our knowledge, the first systematic investigation into bonus incentive effects in nonroutine analytical and collaboratively solved team tasks. To discuss the external validity of our results, we consider it useful to draw on the SANS conditions introduced in List (2020): selection, attrition, naturalness, and scaling.[60] Our two main samples reported in this paper consist of actual ETR customers as well as students, who conceivably differ along several dimensions.[61] As our documented treatment effects carry over to participants from both samples, this seems to indicate that selection is

[58] As HR experts in the survey could have believed that improving teams became substantially faster whereas declining teams became 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, $p = .25$).

[59] In addition to HR experts, in app. sec. A.14, we describe a survey with two samples: a hand-curated list of academics working in personnel economics and respondents from the Economic Science Association's "ESA discuss," a mailing list for academic experimental economists. We asked both samples whether they believed that incentives influence performance in nonroutine analytical team tasks. Over 80% reported 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 toward certainty.

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

[61] As we do not collect background information about customers apart from age, we can assume only that not all ETR participants are university educated (and are 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).

not a primary concern. Additionally, university students are likely (on average) similar to workers in many nonroutine, analytical team work environments, as these frequently require higher levels of education. We also do not consider attrition to be a major concern, as none of the participants opted out from our framed field experiment and participants were unaware of being studied in the natural field experiment (and hence selective attrition could not occur in the latter either).

In terms of scaling, it is worth noting that stakes in our setting are substantially lower than typical bonuses paid in firms. On the other hand, our results in tables 8 and A.3 do not show a significant interaction between incentives and team size, suggesting that, at least locally, the incentive's size is less important. As such, we would expect to observe, if anything, larger effects when applying our interventions in various work environments.[62]

In terms of naturalness, we concede that our task indeed is only one example of a nonroutine 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 nonroutine, analytical team tasks share three features: they (1) are nonroutine, (2) require analytical thinking, and (3) are conducted in teams. Building our experiment around these commonalities ensures that our analysis covers the essence of this class of tasks. This assertion is corroborated by our survey among HR experts, whom we ask about their expectations regarding the efficacy of incentives for team performance either in an escape challenge or in a neutrally framed nonroutine task.

Across both settings, HR experts believe incentives to be similarly effective (see also table A.24). They predict that 41.37% of teams will improve for abstract nonroutine tasks versus 40.38% for escape challenges ($p = .66$, Mann-Whitney test). Furthermore, 21.48% versus 23.33% of teams are predicted to perform worse ($p = .41$, Mann-Whitney test) and 37.15% versus 36.29% similarly ($p = .80$, Mann-Whitney test). While we argue, on the basis of 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 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 generally hold for the choice of an occupation. However,

[62] 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.

many employees working on nonroutine 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 nonroutine tasks in their job.[63] Naturally, self-selection into work environments with nonroutine tasks will likely become less important as current labor market trends continue, with many jobs expected to transform and include more nonroutine 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, nonroutine analytical team tasks are diverse in nature. Intrinsic motivation to perform these tasks (e.g., in business or academia) may stem not only from making progress in and eventually completing them but also from the 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 the lack of such goals reduces external validity, as the effectiveness of incentives may hinge on workers' motivation. However, since we find that incentives increase performance for both people who value performing the task (customer sample) and those being assigned to complete it (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, benefits of which should extend to teams performing tasks with greater goals. Finally, we consider our finding as broadly applicable, as many workers perform nonroutine tasks in occupations that do not necessarily serve greater goals.

Third, one could argue that in some environments, more than one single solution to a complex problem may exist, while in our setting there is only one. We agree that some nonroutine 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 are multiple ways of arriving at that

[63] An intrinsic desire for being able to perform nonroutine analytical jobs has been long recognized and leveraged by recruiters. One notable example is some of Google's recruiting campaigns, which featured signs placed at Harvard Square and across Silicon Valley. These signs were not initially revealed to be associated with Google but instead challenged passersby 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 nonroutine tasks (https://www.eseibusinessschool.com/experimental-escape-room-recruitment-event-esei-tradler/).

solution both in the workplace as well as in our setting. More specifically, we think of incentives as means to motivate the worker to produce the best possible solution in a given amount of time (by identifying the main problems to be solved and coming up with a solution). For example, consider a team of IT specialists 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 employer's demands (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 participating teams. Throughout the task, teams may not know whether there exists only one solution to each subproblem or whether picking one out of a number of possible solutions will let them advance in the task.

Fourth, the proximity of our subjects to their team members may alleviate potential free rider concerns typical in regular office settings. In the absence of free riding, we could thus estimate inflated incentive effects. However, given that the task requires mainly cognitive effort, the observability of coworkers' effort provision is limited in our setting as well. Furthermore, if the utility from completing the task quickly without contributing was lower than in a comparable work setting, we should observe differences in performance effects among highly intrinsically motivated (customer sample) and presumably less intrinsically motivated teams (student sample). However, the 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 the long-run effects of bonus incentives. Instead, we study the general effectiveness of bonus incentives in collaboratively solved nonroutine analytical team tasks in 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 whether the size of the result applies to other nonroutine tasks and occupational environments.

VI. Conclusion

According to Autor, Levy, and Murnane (2003) and Autor and Price (2013), nonroutine, cognitively demanding, interactive tasks are becoming increasingly 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 nonroutine, cognitively demanding team-based 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. In collaboration with our partner, we implemented 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 within 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 collaboratively solved nonroutine 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, rendering the likelihood of crowding out as per 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 that 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 find a tendency toward reduced independent problem-solving and an increased reliance on hints.[64]

[64] 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 since they reach difficult stages earlier. Second, in our natural field experiment, overall hint-taking behavior is not significantly different across treatments. Third,

Second, in contrast to routine tasks, in which the relationship between effort is often deterministic, nonroutine 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, thus making understanding of the production function a crucial mediator for the efficacy of incentives in nonroutine tasks.[65] 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 additional 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 also highlight important tradeoffs 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 provide only 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 nonroutine team work. A very interesting but particularly challenging question that remains is to empirically find the optimal incentive mechanism for performance in nonroutine analytical team tasks. This requires varying different types of incentives (e.g., tournaments, bonuses) and their extent simultaneously, ideally on a set of nonroutine tasks of different nature. While clearly beyond the scope of this study, it is certainly a very interesting and relevant avenue for future research. Looking beyond the question of incentives, we can further use the setting of a real-life escape

when studying at what point in time teams achieve an intermediate step early in the task and how many hints teams have taken before reaching that step, we observe significantly better performance by teams facing incentives but no significant differences in hint taking (see table A.14).

[65] The latter finding also challenges 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.

game to study other important questions, such as goal setting, nonmonetary 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.

Data Availability

Code replicating the tables and figures in this article can be found in Englmaier et al. (2024) in the Harvard Dataverse, https://doi.org/10.7910/DVN/2BADW3.

References

Allen, E. J., P. M. Dechow, D. G. Pope, and G. Wu. 2017. "Reference-Dependent Preferences: Evidence from Marathon Runners." *Management Sci.* 63 (6): 1657–72.

Amabile, T. M. 1996. *Creativity in Context: Update to the Social Psychology of Creativity.* Boulder, CO: Westview.

Antonakis, J., G. d'Adda, R. Weber, and C. Zehnder. 2021. "Just Words? Just Speeches? On the Economic Value of Charismatic Leadership." *Management Sci.* 68 (9): 6355–7064.

Autor, D. H., and M. J. Handel. 2013. "Putting Tasks to the Test: Human Capital, Job Tasks, and Wages." *J. Labor Econ.* 31 (S1): S59–S96.

Autor, D. H., F. Levy, and R. J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *Q.J.E.* 118 (4): 1279–333.

Autor, D. H., and B. Price. 2013. "The Changing Task Composition of the US Labor Market: An Update of Autor, Levy, and Murnane (2003)." Working paper.

Azmat, G., and N. Iriberri. 2010. "The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment Using High School Students." *J. Public Econ.* 94 (7): 435–52.

Azoulay, P., J. S. Graff Zivin, and G. Manso. 2011. "Incentives and Creativity: Evidence from the Academic Life Sciences." *RAND J. Econ.* 42 (3): 527–54.

Bandiera, O., I. Barankay, and I. Rasul. 2005. "Social Preferences and the Response to Incentives: Evidence from Personnel Data." *Q.J.E.* 120 (3): 917–62.

———. 2013. "Team Incentives: Evidence from a Firm Level Experiment." *J. European Econ. Assoc.* 11 (5): 1079–114.

Bandiera, O., G. Fischer, A. Prat, and E. Ytsma. 2021. "Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments." *A.E.R. Insights* 3 (4): 435–54.

Barankay, I. 2010. "Rankings and Social Tournaments: Evidence from a Field Experiment." Working paper.

———. 2012. "Rank Incentives: Evidence from a Randomized Workplace Experiment." Working paper.

Behrman, J. R., S. W. Parker, P. E. Todd, and K. I. Wolpin. 2015. "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools." *J.P.E.* 123 (2): 325–64.

Bénabou, R., and J. Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Rev. Econ. Studies* 70 (3): 489–520.

———. 2006. "Incentives and Prosocial Behavior." *A.E.R.* 96 (5): 1652–78.

Blanes i Vidal, J., and M. Nossol. 2011. "Tournaments without Prizes: Evidence from Personnel Records." *Management Sci.* 57 (10): 1721–36.
Bradler, C., S. Neckermann, and A. J. Warnke. 2019. "Incentivizing Creativity: A Large-Scale Experiment with Performance Bonuses and Gifts." *J. Labor Econ.* 37 (3): 793–851.
Casner-Lotto, J., and L. Barrington. 2006. "Are They Really Ready to Work? Employers' Perspectives on the Basic Knowledge and Applied Skills of New Entrants to the 21st Century U.S. Workforce." ERIC report.
Cassar, L. 2019. "Job Mission as a Substitute for Monetary Incentives: Benefits and Limits." *Management Sci.* 65 (2): 896–912.
Charness, G., and D. Grieco. 2019. "Creativity and Incentives." *J. European Econ. Assoc.* 17 (2): 454–96.
Churchill, G. A., N. M. Ford, and O. C. Walker. 1993. *Sales Force Management: Planning, Implementation, and Control.* Homewood, IL: Irwin/McGraw-Hill.
Corgnet, B., J. Gómez-Miñambres, and R. Hernán-Gonzalez. 2015. "Goal Setting and Monetary Incentives: When Large Stakes Are Not Enough." *Management Sci.* 61 (12): 2926–44.
Deci, E. L., R. Koestner, and R. M. Ryan. 1999. "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bull.* 125 (6): 627–68.
Delfgaauw, J., and R. Dur. 2010. "Managerial Talent, Motivation, and Self-Selection into Public Management." *J. Public Econ.* 94 (9): 654–60.
Delfgaauw, J., R. Dur, A. Non, and W. Verbeke. 2015. "The Effects of Prize Spread and Noise in Elimination Tournaments: A Natural Field Experiment." *J. Labor Econ.* 33 (3): 521–69.
Delfgaauw, J., R. Dur, and M. Souverijn. 2020. "Team Incentives, Task Assignment, and Performance: A Field Experiment." *Leadership Q.* 31 (3): 101241.
Deming, D., and L. B. Kahn. 2018. "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals." *J. Labor Econ.* 36 (S1): S337–S369.
Deming, D. J. 2017. "The Growing Importance of Social Skills in the Labor Market." *Q.J.E.* 132 (4): 1593–640.
Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *J. European Econ. Assoc.* 9 (3): 522–50.
Duncker, K. 1945. "On Problem-Solving." *Psychological Monographs* 58 (5): i–113.
Eckartz, K., O. Kirchkamp, and D. Schunk. 2012. "How Do Incentives Affect Creativity?" CESifo Working Paper no. 4049, CESifo, Munich.
Ederer, F., and G. Manso. 2013. "Is Pay for Performance Detrimental to Innovation?" *Management Sci.* 59 (7): 1496–513.
Englmaier, F., S. Grimm, D. Grothe, D. Schindler, and S. Schudy. 2021. "The Value of Leadership: Evidence from a Large-Scale Field Experiment." CESifo Working Paper no. 9273, CESifo, Munich.
———. 2024. "Replication Data for: 'The Effect of Incentives in Nonroutine Analytical Team Tasks.'" Harvard Dataverse, https://doi.org/10.7910/DVN/2BADW3.
Englmaier, F., A. Roider, and U. Sunde. 2017. "The Role of Communication of Performance Schemes: Evidence from a Field Experiment." *Management Sci.* 63 (12): 4061–80.
Erat, S., and U. Gneezy. 2016. "Incentives for Creativity." *Experimental Econ.* 19 (2): 269–80.

Erev, I., G. Bornstein, and R. Galili. 1993. "Constructive Intergroup Competition as a Solution to the Free Rider Problem: A Field Experiment." *J. Experimental Soc. Psychology* 29 (6): 463–78.
Fehr, D., G. Fink, and B. K. Jack. 2022. "Poor and Rational: Decision-Making under Scarcity." *J.P.E.* 130 (11): 2862–97.
Fehr, E., A. Klein, and K. M. Schmidt. 2007. "Fairness and Contract Design." *Econometrica* 75 (1): 121–54.
Friebel, G., and M. Giannetti. 2009. "Fighting for Talent: Risk-Taking, Corporate Volatility and Organisation Change." *Econ. J.* 119 (540): 1344–73.
Friebel, G., M. Heinz, M. Krüger, and N. Zubanov. 2017. "Team Incentives and Performance: Evidence from a Retail Chain." *A.E.R.* 107 (8): 2168–203.
Fryer, R. G., S. D. Levitt, J. List, and S. Sadoff. 2022. "Enhancing the Efficacy of Teacher Incentives through Framing: A Field Experiment." *American Econ. J. Econ. Policy* 14 (4): 269–99.
Gächter, S., E. J. Johnson, and A. Herrmann. 2022. "Individual-Level Loss Aversion in Riskless and Risky Choices." *Theory and Decision* 92 (3–4): 599–624.
Gerhart, B., and M. Fang. 2015. "Pay, Intrinsic Motivation, Extrinsic Motivation, Performance, and Creativity in the Workplace: Revisiting Long-Held Beliefs." *Ann. Rev. Org. Psychology and Org. Behavior* 2:489–521.
Gibbs, M., S. Neckermann, and C. Siemroth. 2017. "A Field Experiment in Motivating Employee Ideas." *Rev. Econ. and Statis.* 99 (4): 577–90.
Goldszmidt, A., J. A. List, R. D. Metcalfe, I. Muir, V. K. Smith, and J. Wang. 2020. "The Value of Time in the United States: Estimates from Nationwide Natural Field Experiments." Working Paper no. 28208, NBER, Cambridge, MA.
Gosling, S. D., P. J. Rentfrow, and W. B. Swann. 2003. "A Very Brief Measure of the Big-Five Personality Domains." *J. Res. Personality* 37 (6): 504–28.
Gosnell, G. K., J. A. List, and R. D. Metcalfe. 2020. "The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains." *J.P.E.* 128 (4): 1195–233.
Gough, H. G. 1979. "A Creative Personality Scale for the Adjective Check List." *J. Personality and Soc. Psychology* 37 (8): 1398–405.
Harrison, G. W., and J. A. List. 2004. "Field Experiments." *J. Econ. Literature* 42 (4): 1009–55.
Helmreich, R. L., and J. T. Spence. 1978. "The Work and Family Orientation Questionnaire: An Objective Instrument to Assess Components of Achievement Motivation and Attitudes toward Family and Career." *JSAS Catalog Selected Documents Psychology* 8:35.
Hennessey, B. A., and T. M. Amabile. 2010. "Creativity." *Ann. Rev. Psychology* 61:569–98.
Herweg, F., D. Müller, and P. Weinschenk. 2010. "Binary Payment Schemes: Moral Hazard and Loss Aversion." *A.E.R.* 100 (5): 2451–77.
Hoegl, M., and H. G. Gemuenden. 2001. "Teamwork Quality and the Success of Innovative Projects: A Theoretical Concept and Empirical Evidence." *Org. Sci.* 12 (4): 435–49.
Holmstrom, B. 1982. "Moral Hazard in Teams." *Bell J. Econ.* 13 (2): 324–40.
Holmstrom, B., and P. Milgrom. 1991. "Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *J. Law, Econ., and Org.* 7:24–52.
Holz, J. E., J. A. List, A. Zentner, M. Cardoza, and J. E. Zentner. 2023. "The $100 Million Nudge: Increasing Tax Compliance of Firms Using a Natural Field Experiment." *J. Public Econ.* 218:104779.
Hossain, T., and K. K. Li. 2014. "Crowding Out in the Labor Market: A Prosocial Setting Is Necessary." *Management Sci.* 60 (5): 1148–60.

Hossain, T., and J. A. List. 2012. "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations." *Management Sci.* 58 (12): 2151–67.

Jayaraman, R., D. Ray, and F. de Véricourt. 2016. "Anatomy of a Contract Change." *A.E.R.* 106 (2): 316–58.

Jerald, C. D. 2009. "Defining a 21st Century Education." Alexandria, VA: Center Public Educ., Nat. School Boards Assoc.

Kachelmaier, S. J., B. E. Reichert, and M. G. Williamson. 2008. "Measuring and Motivating Quantity, Creativity, or Both." *J. Accounting Res.* 46 (2): 341–73.

Kleine, M. 2021. "No Eureka! Incentives Hurt Creative Breakthrough Irrespective of the Incentives' Frame." Research Paper no. 21-15, Max Planck Inst. Innovation and Competition, Munich.

Kosfeld, M., S. Neckermann, and X. Yang. 2017. "The Effects of Financial and Recognition Incentives across Work Contexts: The Role of Meaning." *Econ. Inquiry* 55 (1): 237–47.

Kuhn, P. J., and L. Yu. 2021. "Kinks as Goals: Accelerating Commissions and the Performance of Sales Teams." Working Paper no. 28487, NBER, Cambridge, MA.

Larkin, I., and S. Leider. 2012. "Incentive Schemes, Sorting, and Behavioral Biases of Employees: Experimental Evidence." *American Econ. J. Microeconomics* 4 (2): 184–214.

Laske, K., and M. Schroeder. 2017. "Quantity, Quality, and Originality: The Effects of Incentives on Creativity." Working paper.

Lazear, E., and P. Oyer. 2013. "Personnel Economics." In *Handbook of Organizational Economics*, edited by R. Gibbons and J. Roberts, 479–519. Princeton, NJ: Princeton Univ. Press.

Lazear, E. P. 2000. "Performance Pay and Productivity." *A.E.R.* 90 (5): 1346–61.

Levitt, S. D., and S. Neckermann. 2014. "What Field Experiments Have and Have Not Taught Us about Managing Workers." *Oxford Rev. Econ. Policy* 30 (4): 639–57.

List, J. A. 2003. "Does Market Experience Eliminate Market Anomalies?" *Q.J.E.* 118 (1): 41–71.

———. 2004a. "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field." *Q.J.E.* 119 (1): 49–89.

———. 2004b. "Neoclassical Theory versus Prospect Theory: Evidence from the Marketplace." *Econometrica* 72 (2): 615–25.

———. 2006. "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions." *J.P.E.* 114 (1): 1–37.

———. 2020. "Non Est Disputandum de Generalizability? A Glimpse into the External Validity Trial." Working Paper no. 27535, NBER, Cambridge, MA.

List, J. A., J. A. Livingston, and S. Neckermann. 2018. "Do Financial Incentives Crowd Out Intrinsic Motivation to Perform on Standardized Tests?" *Econ. Educ. Rev.* 66:125–36.

List, J. A., A. M. Shaikh, and Y. Xu. 2019. "Multiple Hypothesis Testing in Experimental Economics." *Experimental Econ.* 22 (4): 773–93.

Loehlin, J. C., and A. A. Beaujean. 2016. *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis*. London: Taylor & Francis.

Maniadis, Z., F. Tufano, and J. A. List. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *A.E.R.* 104 (1): 277–90.

McCullers, J. C. 1978. "Issues in Learning and Motivation." In *The Hidden Costs of Reward: New Perspectives on the Psychology of Human Motivation*, edited by M. R. Lepper and D. Greene, 5–18. New York: Psychology.

McGraw, K. O. 1978. "The Detrimental Effects of Reward on Performance: A Literature Review and a Prediction Model." In *The Hidden Costs of Reward: New Perspectives on the Psychology of Human Motivation*, edited by M. R. Lepper and D. Green, 33–60. New York: Psychology..

Miller, G., and K. Babiarz. 2014. "Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs." In *Encyclopedia of Health Economics*, edited by A. J. Culyer, 457–66. San Diego: Elsevier.

Moynahan, J. K. 1980. *Designing an Effective Sales Compensation Program*. New York: Amacom.

Mujcic, R., and P. Frijters. 2013. "Economic Choices and Status: Measuring Preferences for Income Rank." *Oxford Econ. Papers* 65 (1): 47–73.

NACE (National Association of Colleges and Employers). 2015. "Job Outlook 2015." Bethlehem, PA: Nat. Assoc. Colleges and Employers.

Oyer, P. 2000. "A Theory of Sales Quotas with Limited Liability and Rent Sharing." *J. Labor Econ.* 18 (3): 405–26.

Pink, D. 2009. "Dan Pink: The Puzzle of Motivation." https://www.ted.com/talks/dan_pink_on_motivation.

Pink, D. H. 2011. *Drive: The Surprising Truth about What Motivates Us*. New York: Riverhead.

Prendergast, C. 1999. "The Provision of Incentives in Firms." *J. Econ. Literature* 37 (1): 7–63.

Ramm, J., S. Tjotta, and G. Torsvik. 2013. "Incentives and Creativity in Groups." CESifo Working Paper no. 4374, CESifo, Munich.

Shearer, B. 2004. "Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment." *Rev. Econ. Studies* 71 (2): 513–34.

Takahashi, H., J. Shen, and K. Ogawa. 2016. "An Experimental Examination of Compensation Schemes and Level of Effort in Differentiated Tasks." *J. Behavioral and Experimental Econ.* 61:12–19.

Ulbricht, R. 2016. "Optimal Delegated Search with Adverse Selection and Moral Hazard." *Theoretical Econ.* 11 (1): 253–78.