CESifo WORKING PAPERS

9273
2021

August 2021

The Value of Leadership: Evidence from a Large-Scale Field Experiment

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

CESifo

Impressum:

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

CESifo *Working Paper No. 9273*

The Value of Leadership: Evidence from a Large-Scale Field Experiment

Abstract

Companies increasingly shift from hierarchical to team-based organizational structures. Scholars and practitioners alike have emphasized the potential of leadership to foster performance in these settings. However, the causal impact of leadership is difficult to identify, as in agile and cross-functional teams leadership is often determined endogenously. This study exploits a unique opportunity to uncover the value of leadership in a non-routine task performed by teams with flat hierarchies. In a large-scale natural field experiment (>1200 participants in 280 teams), we randomly encourage teams to select a leader before performing a complex task. The leadership encouragement increases the fraction of teams solving the task within the given time limit by about 25% and teams' remaining times by roughly 75%. Choosing a leader not only improves performance time-wise, but also team organization, without reducing the originality of solutions. Hence, leadership encouragements can serve as a cost-effective tool to foster team performance.

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

Keywords: teamwork, leadership, non-routine analytical task, complex problem-solving, flat hierarchies.

Florian Englmaier
Department of Economics
University of Munich / Germany
florian.englmaier@econ.lmu.de

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

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

David Schindler
Department of Economics
Tilburg University / Netherlands
d.schindler@tilburguniversity.edu

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

August 25, 2021

We thank John Antonakis, Lea Cassar, David Cooper, Robert Dur, Uri Gneezy, Steve Leider, Steven Levitt, John List, and participants of AFE 2019 (Chicago), HBS Strategy Seminar, MIT Sloan Organizational Economics Lunch, the 4th Lectures in Economics of Organization and Human Ressources (LEOH, Hamburg), Behavioral Personnel EconomicsWorkshop (Santiago de Chile), Field Days 2018 (University of Rotterdam), 4th CRC/TR-190 Conference (Ohlstadt),Workshop on Motivation and Incentives 2018 (University of Stavanger), Advances with Field Experimentes 2019 (Chicago) and seminar participants at MPI Bonn, University of Cologne, Copenhagen Business School, University of Innsbruck, INSEAD, UC3M (Madrid), Prague V˘SE, University of Regensburg, and University of Stirling for helpful thoughts and comments. Further, we thank Silvia Fernandez Castro, Simon Klein and Aloysius Widmann for excellent research assistance. Dominik Grothe acknowledges funding by the Elite Network of Bavaria through IDK Evidence Based Economics. Further, financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) is gratefully acknowledged. This study was pre-registered at the AEA registry (AEARCTR-0002570).

1 Introduction

Competition leads modern firms to flatten hierarchies (Guadalupe and Wulf, 2010), thereby shifting to team-based organizational structures, in which agile and cross-functional teams are confronted with complex, and non-routine analytical tasks (see also Autor et al., 2003; Autor and Price, 2013). This organizational change has important implications for leadership. First, in agile and cross-functional teams, multiple individuals share responsibilities and challenges, rendering the role of leaders ambiguous. Second, cross-functional teams with flat hierarchies often face complex tasks that require team members to exert cognitive effort, stay motivated and work in a coordinated manner. Thus, teams may not only benefit from leaders acting as coaches (Hackman and Wageman, 2005; Morgeson, 2005), modeling or displaying affect (Kaplan et al., 2014; Pirola-Merlo et al., 2002), and managing team boundaries (Druskat and Wheeler, 2003), but also from leaders who explicitly motivate (see, e.g., House, 1976; Bass, 1998, 1999; Howell and Avolio, 1993) and coordinate (see, e.g., Bass, 1990; House et al., 1999) their team members.

While leadership has been attributed importance in business, management, economics, and politics (Antonakis et al., 2019), determining its actual value for teams performing non-routine tasks is particularly challenging. Cross-functional teams are composed of individuals operating on the same hierarchy level such that leadership is often determined endogenously. Consequently, causal estimates of the efficacy of endogenous leadership are largely missing.[1] This study exploits a unique opportunity to uncover the causal effects of the endogenous choice of leaders for team performance in a non-routine team task. To overcome problems of endogeneity and identify causal effects of endogenous leadership, we encourage randomly selected teams to choose a leader before teamwork begins in a pre-registered natural field experiment with 281 teams (consisting of 1,273 participants).

We focus on team performance in a real-life escape room setting, in which teams have to solve a series of cognitively demanding tasks to succeed. Real-life escape rooms encompass important elements encountered in many other non-routine, analytical, and interactive team tasks and are nowadays also used to recruit high-skilled workers as well as

[1] See, for example, the meta-analysis on shared leadership by Nicolaides et al. (2014, p. 936), in which none of the studies included used a true experimental design with randomized control conditions in the field.

to asses and improve individuals' teamwork ability and leadership skills.[2] Escape rooms provide a unique environment to study the value of leadership in non-routine tasks. First, teams need to collect and recombine information, jointly form and test hypotheses, and solve cognitively demanding tasks that require thinking outside the box (see also Englmaier et al., 2018). Second, akin to cross-functional and agile teams, teams performing the task act in flat hierarchies that allow for an endogenous choice of a leader. Third, teams encounter problems that are novel and challenging for them, but kept identical across teams and thus comparable from a performance evaluation perspective. Thus, the setting offers an objective and comparable measure of team performance (teams' likelihood and speed of task completion). Finally, the escape room setting allows us to randomly assign experimental treatment conditions to a large number of teams that are unaware of taking part in an experiment and thus to causally identify the value of leadership in non-routine tasks.

We conduct our natural field experiment (Harrison and List, 2004) in collaboration with the escape room provider *ExitTheRoom*, who allowed us to assign their regular customer teams to two main conditions: *Control* and *Leadership*. The only difference between the two conditions is that in the *Leadership* condition, teams are explicitly asked to select a leader before working on the task while in *Control* they are not. The *Leadership* condition thus emphasizes the positive role of leadership before teamwork starts, but does not enforce the choice of a leader. This simple variation allows for the identification of the value of leadership encouragements in complex teamwork as well as to estimate how choosing a leader affects team performance.

We find a substantial positive effect of *Leadership* on team performance. Treated teams are significantly more likely to complete the task, and complete it considerably faster. The share of teams completing the task within 60 minutes increases from 44% in *Control* to 63% in the *Leadership* condition and teams' average remaining time increases by about 75% (from 3m10s in *Control* to 5m29s in *Leadership*). To delve into potential mechanisms behind the leadership encouragement, we study how different framings of the leader's role (to motivate or to coordinate) within our *Leadership* condition and teams' decision to choose a leader (after being encouraged to do so) affect team performance and team organization. Our results reveal that both framings of the *Leadership* treatment

[2]See e.g. `https://dobetter.esade.edu/en/escape-rooms-business?`, `https://www.eseibusinessschool.com/experimental-escape-room-recruitment-event-esei-tradler/`, and `https://theescapegame.com/virtual-team-building/` (last accessed: June 12, 2021).

yield similarly positive effects on team performance and further, that team performance is significantly better among teams that chose a leader. Findings from two-stage least squares (2SLS) regressions, in which we instrument leader choice by the treatment condition, confirm the efficacy of choosing a leader and indicate that choosing a leader also alters team organization. In teams with leaders, team members tend to be more likely to acquire information individually and less likely to stand together in order to jointly reflect on subtasks. Hence, leadership seems to increase decentralized information acquisition and problem solving. As leadership changes team organization and results in performance increases, it likely improved coordination among team members. This latter interpretation seems is also reflected in teams' perceptions of coordination, which we were allowed to elicited after task performance as part of a short customer survey.

In addition, our setting allows us to consider potential impacts of *Leadership* on the originality of solutions. During the escape room task, teams have the possibility to seek external help by asking for up to five hints if they are stuck. Interpreting the number of hints taken as an inverse measure of teams' willingness to provide original solutions (see also Englmaier et al., 2018), we find that *Leadership* does not decrease originality of solutions nor does it lead to requesting external help earlier.

Taken together, these findings contribute to two strands of the literature. First, our study substantially advances earlier research on the causal effects of leadership. We provide first field evidence on the causal effect of leadership encouragements in teams that may endogenously choose a leader when performing a non-routine task. In contrast to important earlier work that has studied the causal effects of leadership in the field, we focus on the value of choosing a leader instead of comparing different leadership styles (see, e.g., Antonakis et al., 2019; Meslec et al., 2020) or adding different motivational components to leadership speeches (Kvaløy et al., 2015). Further, we focus on teamwork in a non-routine analytical task rather than on individual performance in routine tasks (Kvaløy et al., 2015; Antonakis et al., 2019; Meslec et al., 2020). Related to a large body of laboratory experimental evidence on the positive effects of leadership on coordination (e.g., Weber et al., 2001, 2004; Cooper, 2006; Brandts and Cooper, 2007; Brandts et al., 2007; Cartwright et al., 2013; Sahin et al., 2015; Brandts et al., 2015; Cooper et al., 2020), we further show that leadership can alter team organization and improve (perceived) coordination among team members also in more complex environments.

Second, our study highlights leadership as an important determinant of team performance in non-routine analytical and interpersonal tasks. These tasks have gained substantially in relative importance in the last decades and may gain even more relevance in the age of automation and digitization (Autor et al., 2003; Autor and Price, 2013).[3] Previous work in this domain has focused on the role of monetary incentives for idea creation and team performance (see, e.g., Gibbs et al., 2017; Englmaier et al., 2018, 2021) and found positive incentive effects. Most closely related to our setting, Englmaier et al. (2018) study the effect of offering a bonus for finishing an escape room task faster and found that the bonus increased teams' remaining times on average by a factor of 1.5 and the fraction of teams completing the task by about 10 percentage points. Our leadership encouragement achieves comparable performance improvements. Thus, we identify a substantial value of leadership encouragements for team-performance in non-routine tasks.

Finally, our findings have important implications for practitioners. We show that simply asking teams with flat hierarchies to choose a leader substantially improves performance without impeding on the team's willingness to provide original solutions. In comparison to monetary incentives, such leadership encouragements thus appear as a cost-effective tool to foster team performance. We find that leadership may help to efficiently delegate individual sub-tasks without hampering the teams' ability to efficiently master the challenge they face. Hence, companies may substantially benefit from emphasizing the role of leadership to fostering joint production in agile and cross-functional teams before teamwork begins.

The rest of this manuscript is structured as follows. Section 2 describes our experimental design, measurements and procedures in more detail. We provide results from the experiment in Section 3. Section 4 investigates potential mechanisms, and Section 5 concludes.

[3]These tasks include activities that involve cognitive rather than physical effort, are interpersonal, and involve the forming and testing of hypotheses. More broadly, they also include forms of creative production (see e.g. Ramm et al., 2013; Bradler et al., 2014; Charness and Grieco, 2018; Gibbs et al., 2017; Laske and Schroeder, 2016).

2 Experimental design

2.1 The field setting

We collaborate with *ExitTheRoom* (ETR), a provider of real-life escape rooms.[4] In escape rooms, teams of customers are confronted with a cognitively demanding team challenge, of a non-routine, and interactive nature. The goal is to complete the team challenge within a limited amount of time (60 minutes), and the challenge is composed of a series of quests that ultimately yield a final code to solve the task and succeed. Escape rooms have become increasingly popular over the last years, with more than 2,000 providers in the United States alone, and numerous more in many cities across the globe. Escape room tasks are embedded in a story, for example teams are asked to find a cure for a disease, defuse a bomb, or simply escape from a venue. To complete the task, teams have to search for clues, combine the collected information, and think outside the box. Teams also often need to make unusual use of objects and develop and exchange innovative ideas to arrive at the solution.[5] If a team manages to succeed before the 60 minutes expire, they win; if time runs out before the team solves all quests, they lose.

We conducted our experiments at the facilities of ETR in Munich, Germany. The location offers three rooms with different themes and background stories.[6] Teams face a time limit of 60 minutes and the remaining time is displayed at all times in the rooms. If a team gets stuck, they can request hints via a walkie-talkie from ETR staff. As they can only ask for up to five hints in total, a team needs to state explicitly that they require help. The hints provided by staff never state the direct solution, but only provide vague clues

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

[5] Englmaier et al. (2018, pp. 6-7) provide an example of a typical sub-task in a real-life escape room to illustrate the nature of the task in more detail. We present this example here as well, as our partner asked us not to reveal actual content. In the fictitious setting, a team has found several objects in a room, among them an unlocked 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 team is searching for 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. Similar tasks may further include finding hidden information in pictures, constructing a flashlight out of several parts, or identifying and solving rebus (word picture) puzzles (see also Erat and Gneezy, 2016; Kachelmaier et al., 2008).

[6] In *Madness*, teams need to find the correct code to open a door to escape (ironically) before a mad researcher experiments on them. In *The Bomb*, a bomb and a code to defuse it has to be found. *Zombie Apocalypse* requires teams to find the correct mix of liquids before time runs out (the anti-Zombie potion).

regarding the next required step and teams are informed about their remaining times and hints on large screens on site.

2.2 Experimental treatments and procedures

We pre-registered the experimental design with the AEA registry (AEARCTR-0002570) and conducted our experiment at *ExitTheRoom* between January and March 2018 during their regular opening hours from Monday to Thursday. The 1,273 participants in 281 teams were all regular customers of our collaboration partner. Teams book specific time slots through ETR's website, usually several days in advance. Upon arrival, staff welcomed the teams and teams signed ETR's terms and conditions, including their data privacy policy. Then, ETR staff gave a standardized introduction including the narrative of the booked event and the general rules at ETR. Finally, the staff guided teams to their room. After performing the task, teams participated in a short customer survey.

We implemented two main experimental variations, which we randomized on a daily level.[7] In treatment *Control* (95 teams), staff welcomed teams without further intervention. In treatment *Leadership*, ETR staff in addition highlighted the importance of leadership to succeed in the task, and encouraged teams to select a leader according to a short, standardized script (see below). To more closely investigate the effects of different types of leadership, treatment *Leadership* contained two sub-treatments: *Motivation* (95 teams) and *Coordination* (91 teams). Teams were encouraged to decide on a leader in both sub-treatments, but the conditions differently stressed the role of a leader, as the script used for the instructions shows:

> "One piece of advice before you begin: a good team needs a good **leader**. Past experience has shown that less successful teams often wanted to have been better **led**. Thus, decide on someone of you, who takes over the **leading** role and consistently *motivates / coordinates* the team."[8]

Besides the *differences in instructions* reproduced above, the two sub-treatments were identical. As our main interest lies in establishing the effect of leadership relative to the

[7] In 12 out of 281 cases, ETR staff did not implement the treatment correctly (either by not encouraging leadership at all or by stimulating the wrong leadership function). In the appendix, we show that our main conclusions do not hinge on the inclusion of these observations.

[8] Bold printed text highlights that leadership was saliently encouraged in the message. Text in *italics* indicates treatment differences in terms of the framing of the leader's function. In treatment *Motivation* ETR staff mentions the word 'motivates', while in treatment *Coordination* they use the word 'coordinates'.

Control condition, we pool the data for the main analyses and utilize both sub-treatments when discussing mechanisms, to show that framing the leader's role according to a specific function (motivation or coordination) does not affect team performance differentially.

2.3 Outcome measures and sample characteristics

In all conditions, we collected observable information related to team performance and team characteristics. These include the time needed to complete the task, the number and timing of requested hints, team size, gender and age composition of the team, the language a team spoke (German or English), experience with escape rooms, and whether the customers came as a private group or were part of a company team-building event.[9] Additionally, as a proxy for the teams' propensity to have someone taking the lead, we collected information about whether one team member actively took possession of the hand-held walkie-talkie and recorded whether the teams explicitly chose a leader before entering their room. While teams were working on the task, our research assistants watched the live CCTV (no audio) and took notes on whether team members searched for information individually (as opposed to jointly), and whether teams were spending much time standing together (versus spread out across the room) on a 5-point Likert scale (from 1 = 'not at all' to 5 = 'a lot').[10] Table 1 compares all pre-determined variables across samples and highlights that our sample is balanced in terms of teams' observable characteristics. To account for minor differences in observable characteristics we provide both non-parametric treatment comparisons and regression analyses that control for additional covariates. Our primary outcome variable in these analyses is team performance, which we measure by i) whether or not teams completed the task in 60 minutes, and ii) the time remaining upon completion. We estimate the causal effect of encouraging leadership on these objective performance measures by comparing the *Leadership* treatment with the *Control* condition. Further outcomes include the number of hints taken as well as responses to a short (five-question) customer survey teams completed after experi-

[9] All these variables were either directly observable to us or were recorded as part of the standard questions ETR's staff asked customers, apart from age. In order to preserve the main characteristics of a natural field experiment and to avoid any study-awareness, we did not ask for the age of participants. Instead, our research assistants estimated each persons age based on their appearance to be either between 18 and 25 years, 26 and 35 years, 36 and 50 years, or above 50 years.

[10] For reasons of data protection, ETR does not keep any video recordings of the team challenge.

Table 1: Sample size and team characteristics

	Control (n=95)	*Leadership* (n=186)
Group Size	4.41 (1.12) [2,7]	4.59 (0.92) [2,6]
Experience with Escape Rooms	0.76 (0.43) [0,1]	0.72 (0.45) [0,1]
Private Event	0.76 (0.43) [0,1]	0.73 (0.44) [0,1]
Share of Male Participants	0.54 (0.29) [0,1]	0.52 (0.30) [0,1]
Median Age	32.43 (8.91) [21.5,55]	32.99 (8.21) [21.5,55]
German-Speaking	0.84 (0.37) [0,1]	0.93 (0.26) [0,1]
One Team Member Actively Took Walkie-Talkie	0.69 (0.46) [0,1]	0.76 (0.43) [0,1]

Notes: For all variables, we report means on the group level. Experience with Escape Rooms is a dummy defined as teams having at least one member with escape game experience. Private Event is a dummy, where professional or team-building events are coded as 0. Median age is constructed as the median of all team members' estimated age, where each individual team member's age is defined as the midpoint of the age categories: 18-25 (21.5), 26-35 (30.5), 36-50 (43), 51+ (assumed to be 55). Standard deviations and minimum and maximum values in parantheses; (std. err.) [min,max]. Stars indicate significant differences to Control applying the procedure for multiple hypothesis testing proposed by List et al. (2019) with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

encing the escape room. This survey included questions on overall satisfaction with the team challenge, the value for money, exerted effort level, and how teams perceived coordination and motivation in the team. All questions were answered on a 8-point Likert scale.

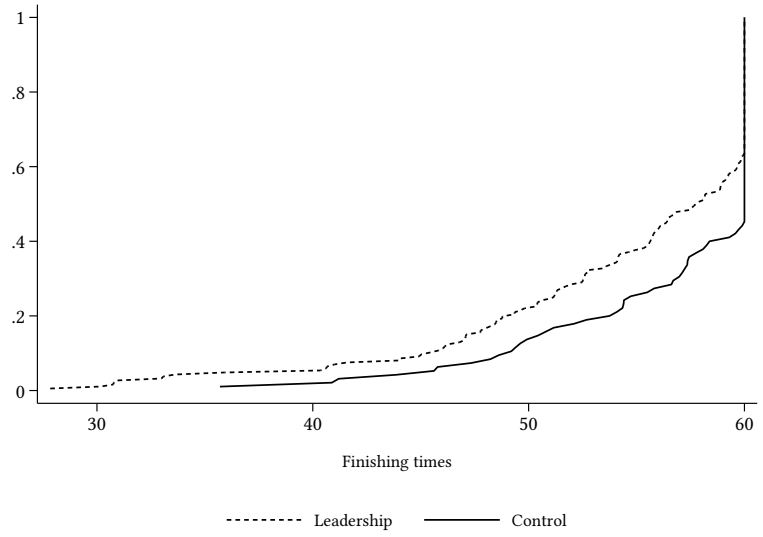
3 Results

3.1 Team performance

Figure 1 shows the cumulative distribution functions of finishing times across conditions. Teams in the *Leadership* treatment conditions perform clearly better than those in our *Control* condition. Specifically, 63% of teams finish the task within the time limit of 60 minutes in *Leadership*, whereas only around 44% do so in *Control* (Pearson χ^2 test: $p < 0.01$). In addition to being more likely to complete the task, teams that were encouraged to choose a leader also solve the task faster (Mann-Whitney test: $p < 0.01$).

These non-parametric results are confirmed by a series of Probit regressions, in which we step-wise introduce additional control variables. To account for differences in the task teams face, all specifications include room fixed effects. In Column (1) of Table 2, we estimate the average marginal effect of *Leadership* on the probability to complete the task within 60 minutes without the inclusion of any additional covariates. In Column (2), we add observable team characteristics (as described in Table 1). To account for potentially idiosyncratic behavior by ETR staff who delivered the general instructions and the leadership encouragement, we employ staff member (including our own research

Figure 1: CDFs of finishing time



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

assistants) fixed effects in Column (3). Finally, in Column (4) we include fixed effects to control for the week of the year and the day of the week. We cluster standard errors at the daily level, which is also the level of random treatment assignment. In all specifications, we find that *Leadership* significantly increases teams' probability to succeed within 60 minutes. The estimated average marginal effect amounts to an increase of 11 percentage points as compared to *Control*, implying a relative increase in the fraction of successful teams of about 25% as compared to the *Control* condition.

The cumulative distribution functions of finishing times in *Leadership* and *Control* (see Figure 1) indicates that teams in our treatment condition *Leadership* solve the task not only more frequently within 60 minutes but also substantially faster. The CDF of finishing times in *Control* stochastically dominates the CDF of *Leadership*, and the data skew towards the end, and are very flat in the left tail. Further, finishing times are censored at 60 minutes. To avoid an underestimation of the treatment effect and take censoring into account, we estimate the effect of *Leadership* on finishing times using a series of Tobit (instead of OLS) regressions and, add additional controls in a step-wise fashion (analogously to the Probit models presented earlier). Table 2 reveals a statistically significant and sizable reduction of finishing times in *Leadership* in all four specifications. Teams

Table 2: Team performance (completion and finishing time)

	Completed within 60 minutes				Finishing time			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Leadership	0.137***	0.137***	0.125**	0.108**	-3.175***	-3.037***	-2.773**	-2.551**
	(0.045)	(0.047)	(0.058)	(0.043)	(0.912)	(0.873)	(1.137)	(1.253)
Mean in Control	0.442	0.442	0.442	0.442	56.814	56.814	56.814	56.814
Observations	281	281	281	281	281	281	281	281
Team Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Staff FE	No	No	Yes	Yes	No	No	Yes	Yes
Weekday and Week FE	No	No	No	Yes	No	No	No	Yes

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

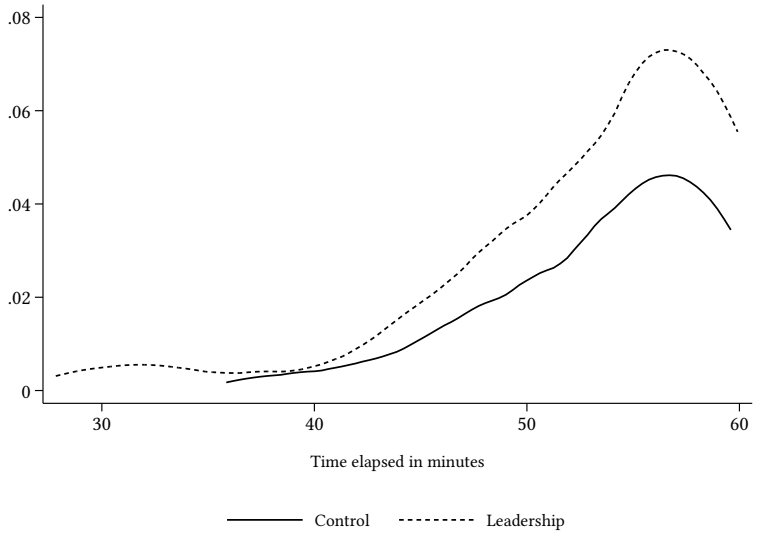
are on average two and a half minutes faster, which is equivalent to an increase of about 75% of teams' remaining times.

Finally, in Figure 2 we provide results from a hazard model (survival analyses) in which finishing the task is considered the "hazard". Figure 2 illustrates hazard rates of completing the task, conditional on not yet having it completed, separately for both conditions. The figure shows that for both treatments the hazard rate is increasing over time (until shortly before the end). Teams' likelihood of completion naturally increases the more time they have invested, but decreases in the last five minutes, conditional on the fact that they have not yet found the solution. Most importantly, the figure reveals a striking absolute difference in the hazard rates between *Leadership* and *Control*. At any given point in time, teams that were encouraged to select a leader face a higher chance of eventually completing the task successfully. The gap between hazard rates in *Leadership* and *Control* starts to widen around the 40-45 minute mark, indicating that leadership most likely affected teams below the top performers and more so teams with intermediate finishing times, while we do not find that leadership substantially improved performance of teams at the lower end of the performance distribution.

3.2 Robustness

To explore the robustness of our estimates from the two previous sections, we perform an (even more conservative) randomization inference exercise (Young, 2019). In our data, we randomly assign each team to either condition, independently of the condition teams were actually assigned to. We then estimate the effect of *Leadership* for this counter-

Figure 2: Hazard rates of finishing the task

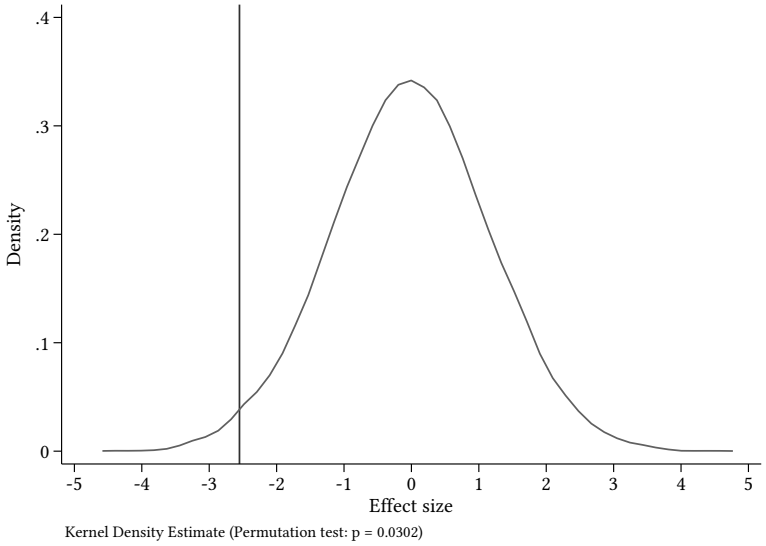


Notes: The figure shows the hazard rates of finishing the task (conditional on not having finished yet) separately for teams we randomly encouraged to select a leader (*Leadership*) and teams in the *Control* condition.

factual. This procedure is repeated 10,000 times, generating a distribution of counterfactual estimates we can compare to our "true" estimate. Figure 3 plots the distributions for teams' finishing times. The Kernel density estimate is centered at 0 and appears normally distributed. The vertical, solid line indicates the observed effects based on the true treatment assignment. As can be seen, the observed effects is "extreme" such that we can confidentially reject the null hypothesis of no effect of our actual treatment ($p-$value$= 0.0302$).

Further robustness analyses are relegated to the Appendix. Appendix Table A.1 repeats the specifications from Table 2 but excludes the 12 observations, where ETR staff implemented the wrong treatment. Our conclusions remain unaffected. Appendix Table A.2 shows results from linear probability models (instead of the earlier used Probit regressions) to estimate the probability of our treatment on a team's success and a generalized linear model with log link to account for the count-like data structure, with finishing times as the dependent variable. The effect of our leadership intervention is of a similar magnitude and significance as reported in Table 2. Further, Appendix Tables A.3 and A.4 provide additional heterogeneity analyses based on observable teams characteristics (see

Figure 3: Randomization inference



Notes: The figure plots the distributions of the effect sizes of Leadership on teams' finishing time using 10,000 repetitions of randomly assigning treatment. The effect size is teams' change in the finishing time, the vertical, solid line indicates the treatment effect observed in the experiment.

also Table 1). Using models with interaction terms, we do not find strong differences in the efficacy of *Leadership* based on underlying team characteristics.[11]

[11]Only one out of the fourteen interaction terms (the interaction with whether a team speaks German in the regression for completing the task within 60 minutes) is negative and statistically significant at the five percent level. The result should however be taken with a grain of salt, as only a small minority of teams does not speak German.

4 Mechanisms

4.1 The framing of leadership functions

As described in Section 2.2, we framed the role of leaders differently in two sub-treatments *Motivation* and *Coordination*. In *Motivation*, we suggested that the group may want to choose a leader who motivates the team, while in sub-treatment *Coordination*, we emphasized that teams may choose a leader to coordinate the team. In Table 3, we estimate the effect of each sub-treatment separately. Our findings show that both sub-treatments are similarly effective. The average marginal effect of *Motivation* (*Coordination*) in our Probit specifications in Column (1) amounts to 13.4 (9.3) percentage points, and finishing times are also significantly reduced in both sub-treatments. A post-estimation Wald test cannot reject the equality of coefficients in either case. Hence, leadership encouragement per se rather than making participants aware of the importance of certain leadership functions is responsible for the observed performance increase.[12]

Table 3: Effects of motivation and coordination on team performance

	Completed within 60 minutes (1)	Finishing time (2)
Motivation	0.134**	-3.482**
	(0.053)	(1.588)
Coordination	0.093**	-2.015*
	(0.042)	(1.198)
Mean in Control	0.442	56.814
Observations	281	281
Team Controls	Yes	Yes
Staff FE	Yes	Yes
Weekday and Week FE	Yes	Yes
Motivation = Coordination	p = 0.316	p = 0.201

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

4.2 Choosing a leader

Next, we investigate whether those teams who actually chose a leader also perform better. Around 50 percent of teams encouraged to choose a leader do so before working on

[12] As the treatment difference between *Coordination* and *Motivation* was rather subtle, it is an interesting avenue for future research to investigate whether a stronger and more salient framing of these functions is able to expand on the overall effect of leadership we detected.

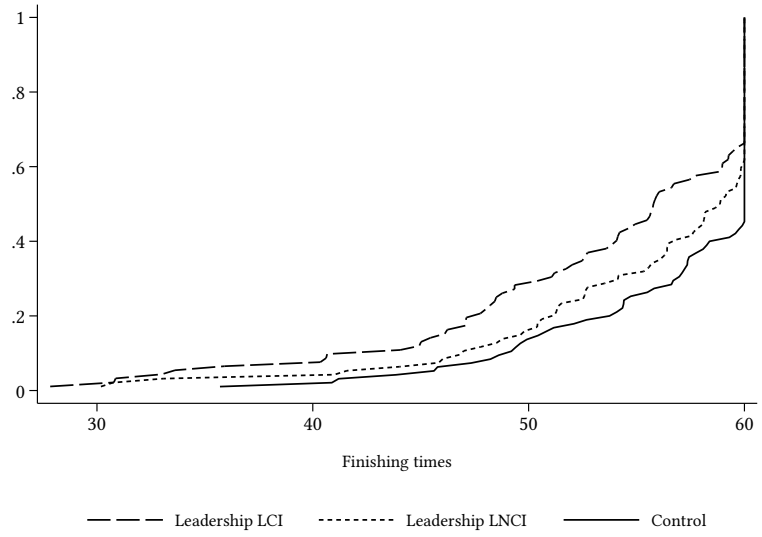
the task, whereas we did not observe a single team explicitly choosing a leader in *Control* before teamwork began. Regression analyses in Appendix Table A.5 further indicate that the immediate choice of a leader does not relate systematically to observable team characteristics.[13] As choosing a leader is equally likely in both sub-treatments (see Appendix Table A.5, Column (2)), we again focus on our main treatment condition *Leadership*. Figure 4 shows the cumulative distribution functions of finishing times in *Leadership* depending on whether a leader was chosen immediately (LCI) or not chosen immediately (LNCI) as well as finishing times of teams in *Control*. The figure illustrates two interesting findings: First, independent of whether teams immediately decided on a leader or not, team performance improves both on the intensive margin (Mann-Whitney: LCI vs. Control, $p < 0.01$; LNCI vs. Control, $p < 0.10$,) and the extensive margin (Pearson χ^2: LCI vs. Control, $p < 0.01$; LNCI vs. Control, $p < 0.05$). Second, teams that were encouraged to choose a leader and chose a leader immediately (LCI) tend to outperform teams that were encouraged but did not chose a leader immediately (LNCI) at the intensive margin (Mann-Whitney: LCI vs. LNCI, p $= 0.09$), but less so at the extensive margin (Pearson χ^2: LCI vs. LNCI, p $= 0.51$).[14]

To analyze whether teams immediately choosing a leader were more successful, we follow the procedure recommended in Angrist and Pischke (2008, p. 142) and employ a two-stage approach. In a first step, we predict the probability of immediately choosing a leader using a Probit model (accounting for the same fixed effects and control variables as in our previous specifications). In a second step, we use these non-linear fitted values as instruments and estimate their impact on team performance. Table 4 presents results from OLS and 2SLS regressions for comparison. Panel A reports the intention-to-treat estimates of regressing a dummy on whether a team completed the task within 60 minutes (Column (1)) or the finishing time (Column (2)) on being assigned to the *Leadership* condition. Panel B contains the two-stage least square results of the second stage. Further,

[13] Similarly, as shown in Appendix Table A.5 Column (3), observable team characteristics have limited predicted power for the chosen leader's gender (less male teams, older teams, and non-German speaking teams are more likely to select a female leader). Further note that our design was not tailored to measure the impact of different leadership characteristics (as these are endogenously determined in our setting) and as we have only very limited knowledge about the leaders observable characteristics (research assistants only took note of the leader's gender). We thus consider the discussion on who is chosen as a leader an interesting question for future research.

[14] To avoid study awareness and preserve the nature of a natural field experiment, we did not ask teams at any later stage whether they chose a leader. Hence, LNCI and *Control* teams may be composed of teams who never chose a leader and teams who chose a leader at a later stage while performing the task.

Figure 4: CDFs of finishing times



Notes: The figure shows the cumulative distribution of finishing times of treated teams that chose a leader immediately (*Leadership* LCI), that were assigned to treatment, but did not choose a leader immediately (*Leadership* LNCI), and teams that were assigned to *Control*.

the table displays the means of dependent variables in *Control*, and a Kleibergen-Paap Wald F statistic of 604.7, indicating that the instrument appears relevant. Column (1) shows that the OLS intention-to-treat estimate in Panel A amounts to 0.112 while the coefficient for the instrumented choice of a leader in Panel B is 0.145. Further, results in Column (2) indicate that the coefficient of immediately choosing a leader in Panel B is larger than the intention-to-treat estimate in Panel A, indicating that teams choosing a leader immediately are indeed more successful and solve the task substantially faster.

4.3 Leaders and their impact

Although our experiment was mainly designed to test the causal impact of a simple leadership encouragement on team performance, we collected additional measures that allow us to discuss how the performance increase through leadership potentially comes about. Most importantly, our research assistants took notes on teams' tendency to stand together and to search individually for information. Acquiring information individually may be beneficial if the team is well organized and exchanges the collected information, while standing together may indicate joint acquisition or reflection on ideas, which may be less

Table 4: Effects of leadership on team performance

	Completed within 60 minutes (1)	Finishing time (2)
Panel A. OLS (ITT)		
Leadership	0.112**	-1.326*
	(0.048)	(0.744)
Panel B. 2SLS (2nd Stage)		
Chose leader immediately	0.145**	-2.761***
	(0.073)	(1.067)
Mean in Control	0.442	56.814
Observations	281	281
Team Controls	Yes	Yes
Staff FE	Yes	Yes
Weekday and Week FE	Yes	Yes
Kleibergen-Paap Wald F	604.7	604.7

Notes: The table displays coefficients from OLS (Panel A) and 2SLS (Panel B) regressions of whether a team solved the task within 60 minutes or finishing times on our treatment indicator (with Control as base category). For 2SLS we follow the procedure outlined in Angrist and Pischke (2008): In a first step, we predict the probability of immediately choosing a leader using all control variables and fixed effects, as well as our treatment indicator in a Probit model. Then, we use these nonlinear fitted values as instruments in the second stage. All columns include room fixed effects. Each column indicates whether team controls (group size, share of male participants, experience with escape games, median age, language spoken, private versus team-building events, actively taken walkie-talkie), staff, weekday and week fixed effects are included. Standard errors in parentheses are clustered at the daily level, with significance levels * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

relevant when teams are well organized. Table 5 shows estimates from OLS and 2SLS regressions (using to the same approach as in Table 4) for the impact of *Leadership* and of choosing a leader on teams' (standardized) tendency to stand together and searching individually for information. The ITT estimate in Column (1), Panel A shows that being assigned to the *Leadership* condition has a significant negative effect on team members' tendency to stand together, and this effect is even more pronounced when teams chose a leader (Column (1), Panel B). The ITT estimate shown in Column (2), Panel A, further indicates that our *Leadership* encouragement increased teams' propensity to search individually, and even more so, when teams chose a leader. This suggests that our leadership encouragement is effective because it increases teams' tendency to choose a leader and because it changes teams' strategy to acquire and process information, particularly for teams that chose a leader. As, overall, leadership results in a substantial performance increase, teams who changed their strategies to acquire and process information in *Leadership* were likely also better organized. In line with this reasoning, we observe that teams in *Leadership* seem to rate their team coordination by about 0.325 of a standard deviation better than teams in *Control* (see Appendix Table A.6, Column (5), in which we utilize teams' responses to the short customer survey).

Finally, our setting also allows us to study whether leaders affect how much teams rely on external help. Recall that in the task all teams can request up to five hints by

Table 5: Effects of leadership on team organization

	Standing Together (1)	Individual Search (2)
Panel A. OLS (ITT)		
Leadership	-0.220**	0.234**
	(0.107)	(0.106)
Panel B. 2SLS (2nd Stage)		
Chose leader immediately	-0.417**	0.375*
	(0.174)	(0.210)
Observations	279	279
Team Controls	Yes	Yes
Staff FE	Yes	Yes
Weekday and Week FE	Yes	Yes
Kleibergen-Paap Wald F	692.5	692.5

Notes: The table displays coefficients from OLS (panel A) and 2SLS (panel B) regressions of how much teams stand together and search individually on our treatment indicator (with Control as base category). All variables are standardized with mean zero and standard deviation of one. For 2SLS we follow the procedure outlined by Angrist and Pischke (2008): in a first step, we predict the probability of immediately choosing a leader using all control variables and fixed effects, as well as our treatment indicator in a Probit model. Then, we use these nonlinear fitted values as instruments in the second stage. All columns include room fixed effects. Each column indicates whether team controls (group size, share of male participants, experience with escape games, median age, language spoken, private versus team-building events, actively taken walkie-talkie), staff, weekday and week fixed effects are included. Standard errors in parentheses are clustered at the daily level, with significance levels * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

contacting ETR staff using a walkie-talkie if they get stuck. In Table 6, we present regression results regarding the impact of *Leadership* on the number of hints and the timing of requesting these hints. The results in Column (1) report the total number of hints requested as the outcome variable. There is no significant difference between teams in our *Leadership* and *Control* condition. Additionally, the analyses in Columns (2) to (6) suggest that *Leadership* has also a minor influence on the timing of hints. We thus conclude that *Leadership* improves team performance without negatively affecting the originality of provided solutions.

Table 6: Effects of leadership on originality

	Hints (1)	1st Hint (2)	2nd Hint (3)	3rd Hint (4)	4th Hint (5)	5th Hint (6)
Panel A. OLS (ITT)						
Leadership	0.047	0.386	0.614	-0.172	-0.074	-0.159
	(0.146)	(1.455)	(1.425)	(1.160)	(0.589)	(0.275)
Panel B. 2SLS (2nd Stage)						
Chose leader immediately	-0.087	1.077	0.536	0.099	0.317	-0.315
	(0.225)	(2.212)	(1.993)	(1.597)	(0.922)	(0.413)
Mean in Control	3.421	21.175	35.115	47.264	54.518	58.815
Observations	281	268	239	204	141	72
Team Controls	Yes	Yes	Yes	Yes	Yes	Yes
Staff FE	Yes	Yes	Yes	Yes	Yes	Yes
Weekday and Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Kleibergen-Paap Wald F	604.7	604.7	604.7	604.7	604.7	604.7

Notes: The table displays coefficients from OLS (Panel A) and 2SLS (Panel B) regressions of whether a team solved the task within 60 minutes or finishing times on our treatment indicator (with Control as base category). For 2SLS we are following the procedure described by Angrist and Pischke (2008): in a first step, we predict the probability of choosing a leader immediately using all control variables and fixed effects as well as our treatment indicator using a Probit model. Then, we use these nonlinear fitted values as instruments. All columns include room fixed effects. Each column indicates whether team controls (group size, share of male participants, experience with escape games, median age, language spoken, private versus team-building events, actively taken walkie-talkie), staff, weekday and week fixed effects are included. Standard errors in parentheses are clustered at the daily level, with significance levels * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

5 Conclusion

This work exploits the unique opportunity to study the causal effect of leadership in a non-routine analytical team task. Motivated by the recent shift in firm organization (Guadalupe and Wulf, 2010) from vertical to horizontal, team based structures, we investigate whether performance in teams with flat hierarchies can be improved by a simple encouragement to choose a leader before team work begins. We conducted a large scale natural field experiment (Harrison and List, 2004) with more than 280 teams performing an escape room challenge, in which we randomly assigned teams to a *Leadership* encouragement or *Control* condition. We document a substantial and robust positive influence of leadership. Asking teams to decide on a leader improves performance on the extensive as well as on the intensive margin. In the *Leadership* condition, 63% of teams complete the task within the given time limit while only 44% of teams do so in *Control*. Further, teams in *Leadership* complete the task substantially faster. The times remaining until the deadline are about 75% larger. The observed treatment effect was mostly driven by teams immediately following the encouragement to choose a leader, and came hand in hand with a change in team organization. The *Leadership* encouragement increased decentralized information acquisition and problem solving, and improved team organization, without reducing the originality of solutions.

Apart from immediate implications for cost-effective improvements of team performance through leadership encouragements in practice, these findings highlight also many interesting avenues for future research. First, it appears natural to investigate the value of endogenous leadership as compared to an exogenous assignment of leaders. Second, and inspired by the changes in team organization identified in this work, there remain many interesting micro-aspects of leadership to be uncovered. For example, future work may study how leadership alters communication, task allocation, and heterogeneity in team members' effort provision, as well as how particular leadership characteristics may causally affect team performance and team organization in non-routine tasks.[15] Further, building on previous work that has investigated the interaction of monetary incentives and particular leadership functions such as motivational speeches (Kvaløy et al., 2015) or verbal feedback (Manthei et al., 2019), a fruitful avenue for future research lies in studying whether endogenous leadership in flat hierarchies and incentives are substitutes or complements. Finally, following theoretical arguments by Hermalin (1998) and Bolton et al. (2008), it will be interesting to investigate which leadership styles most likely overcome information asymmetries among team members in complex teamwork, and whether it matters that a leader is developing a team's strategy (see also Van den Steen, 2018) and how the leader's legitimacy influences strategy implementation.

[15]For interesting recent contributions in this context, see, e.g., De Paola et al. (2018), Fest et al. (2019), and Dur et al. (forthcoming).

References

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Antonakis, J., d'Adda, G., Weber, R., and Zehnder, C. (2019). Just words? just speeches? on the economic value of charismatic leadership. *NBER Rep. 4*.

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

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

Bass, B. M. (1990). From transactional to transformational leadership: Learning to share the vision. *Organizational Dynamics*, 18(3):19–31.

Bass, B. M. (1998). Transformational leadership: Industry, military, and education impact. *New Jersey: Lawrence Erlbaum Associates*.

Bass, B. M. (1999). Two decades of research and development in transformational leadership. *European Journal of Work and Organizational Psychology*, 8(1):9–32.

Bolton, P., Brunnermeier, M. K., and Veldkamp, L. (2008). Leadership, coordination and mission-driven management. Technical report, National Bureau of Economic Research.

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

Brandts, J. and Cooper, D. J. (2007). It's what you say, not what you pay: An experimental study of manager-employee relationships in overcoming coordination failure. *Journal of the European Economic Association*, 5(6):1223–1268.

Brandts, J., Cooper, D. J., and Fatas, E. (2007). Leadership and overcoming coordination failure with asymmetric costs. *Experimental Economics*, 10(3):269–284.

Brandts, J., Cooper, D. J., and Weber, R. A. (2015). Legitimacy, communication, and leadership in the turnaround game. *Management Science*, 61(11):2627–2645.

Cartwright, E., Gillet, J., and Van Vugt, M. (2013). Leadership by example in the weak-link game. *Economic Inquiry*, 51(4):2028–2043.

Charness, G. and Grieco, D. (2018). Creativity and incentives. *Journal of the European Economic Association*.

Cooper, D. J. (2006). Are experienced managers experts at overcoming coordination failure? *Advances in Economic Analysis & Policy*, 5(2).

Cooper, D. J., Hamman, J. R., and Weber, R. A. (2020). Fool me once: An experiment on credibility and leadership. *The Economic Journal*, 130(631):2105–2133.

De Paola, M., Gioia, F., and Scoppa, V. (2018). Teamwork, leadership and gender. Technical report, IZA Discussion Papers.

Druskat, V. U. and Wheeler, J. V. (2003). Managing from the boundary: The effective leadership of self-managing work teams. *Academy of Management Journal*, 46(4):435–457.

Dur, R., Kvaloy, O., and Schöttner, A. (forthcoming). Labor-market conditions and leadership styles. *Managment Science*.

Englmaier, F., Grimm, S., Schindler, D., and Schudy, S. (2018). The effect of incentives in non-routine analytical team tasks-evidence from a field experiment. *CEPR Discussion Paper 13226*.

Englmaier, F., Grothe, D., Grimm, S., Schindler, D., and Schudy, S. (2021). The efficacy of tournaments for non-routine team tasks. *mimeo*.

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

Fest, S., Kvaloy, O., Nieken, P., and Schöttner, A. (2019). Motivation and incentives in an online labor market.

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

Guadalupe, M. and Wulf, J. (2010). The flattening firm and product market competition: The effect of trade liberalization on corporate hierarchies. *American Economic Journal: Applied Economics*, 2(4):105–27.

Hackman, J. R. and Wageman, R. (2005). A theory of team coaching. *Academy of Management Review*, 30(2):269–287.

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

Hermalin, B. E. (1998). Toward an economic theory of leadership: Leading by example. *American Economic Review*, pages 1188–1206.

House, R. (1976). *A 1976 Theory of Charismatic Leadership*. Working paper series - University of Toronto, Faculty of Management Studies. University of Toronto, Faculty of Management Studies.

House, R. J., Hanges, P., Ruiz-Quintanilla, S., Dorfman, P., Javidan, M., Dickson, M., Mobley, W., Gessner, M., and Arnold, V. (1999). Advances in global leadership.

Howell, J. M. and Avolio, B. J. (1993). Transformational leadership, transactional leadership, locus of control, and support for innovation: Key predictors of consolidated-business-unit performance. *Journal of Applied Psychology*, 78(6):891.

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

Kaplan, S., Cortina, J., Ruark, G., LaPort, K., and Nicolaides, V. (2014). The role of organizational leaders in employee emotion management: A theoretical model. *The Leadership Quarterly*, 25(3):563–580.

Kvaløy, O., Nieken, P., and Schöttner, A. (2015). Hidden benefits of reward: A field experiment on motivation and monetary incentives. *European Economic Review*, 76:188–199.

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

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

Manthei, K., Sliwka, D., and Vogelsang, T. (2019). Talking about performance or paying for it? evidence from a field experiment.

Meslec, N., Curseu, P. L., Fodor, O. C., and Kenda, R. (2020). Effects of charismatic leadership and rewards on individual performance. *The Leadership Quarterly*, page 101423.

Morgeson, F. P. (2005). The external leadership of self-managing teams: intervening in the context of novel and disruptive events. *Journal of Applied Psychology*, 90(3):497.

Nicolaides, V. C., LaPort, K. A., Chen, T. R., Tomassetti, A. J., Weis, E. J., Zaccaro, S. J., and Cortina, J. M. (2014). The shared leadership of teams: A meta-analysis of proximal, distal, and moderating relationships. *The Leadership Quarterly*, 25(5):923–942.

Pirola-Merlo, A., Härtel, C., Mann, L., and Hirst, G. (2002). How leaders influence the impact of affective events on team climate and performance in r&d teams. *The Leadership Quarterly*, 13(5):561–581.

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

Sahin, S. G., Eckel, C., and Komai, M. (2015). An experimental study of leadership institutions in collective action games. *Journal of the Economic Science Association*, 1(1):100–113.

Van den Steen, E. (2018). Strategy and the strategist: How it matters who develops the strategy. *Management Science*, 64(10):4533–4551.

Weber, R., Camerer, C., Rottenstreich, Y., and Knez, M. (2001). The illusion of leadership: Misattribution of cause in coordination games. *Organization Science*, 12(5):582–598.

Weber, R. A., Camerer, C. F., and Knez, M. (2004). Timing and virtual observability in ultimatum bargaining and weak link coordination games. *Experimental Economics*, 7(1):25–48.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2):557–598.

A Appendix

A.1 Additional robustness analyses

In this section, we present results on the robustness of the observed treatment effect. Table A.1 repeats the specifications from Table 2 but excludes the 12 observations, where ETR staff implemented the wrong treatment. The results are very similar. Only one specification (not our preferred one) lacks statistical significance (Column (3)), but the coefficients are all of similar magnitude. Table A.2 reports findings from a linear probability model estimating the impact of *Leadership* on the probability to solve the task and GLM estimations on teams' finishing times.

Table A.1: Team performance (completion and finishing time)

	Completed within 60 minutes				Finishing time			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Leadership	0.127***	0.108**	0.088	0.080*	-3.416***	-2.905***	-2.619**	-2.898**
	(0.046)	(0.052)	(0.065)	(0.046)	(0.836)	(0.881)	(1.211)	(1.156)
Mean in Control	0.447	0.447	0.447	0.447	57.063	57.063	57.063	57.063
Observations	269	269	269	269	269	269	269	269
Team Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Staff FE	No	No	Yes	Yes	No	No	Yes	Yes
Weekday and Week FE	No	No	No	Yes	No	No	No	Yes

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

Table A.2: Team performance (completion and finishing time)

	Completed within 60 minutes				Finishing time			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Leadership	0.143***	0.141***	0.130**	0.112**	-0.025***	-0.023**	-0.020*	-0.022*
	(0.048)	(0.049)	(0.062)	(0.048)	(0.009)	(0.009)	(0.010)	(0.012)
Mean in Control	0.442	0.442	0.442	0.442	4.035	4.035	4.035	4.035
Observations	281	281	281	281	281	281	281	281
Team Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Staff FE	No	No	Yes	Yes	No	No	Yes	Yes
Weekday and Week FE	No	No	No	Yes	No	No	No	Yes

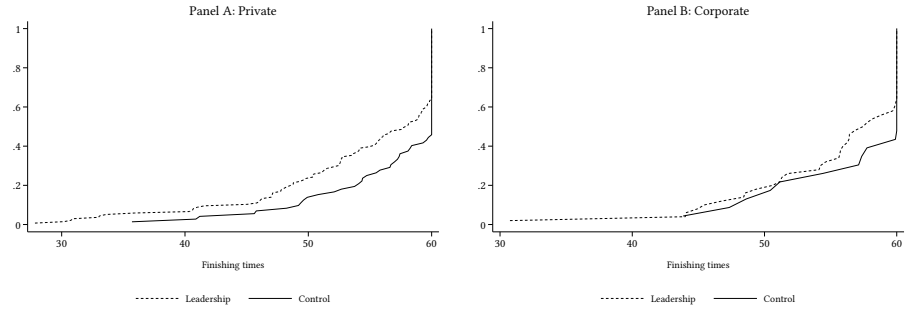
Notes: The table displays coefficients from OLS regressions of whether a team completed the task within 60 minutes (Columns (1) through (4)), and GLM regressions (with log link) of finishing time (Columns (5) through (8)) on our *Leadership* indicator (with *Control* as base category). All columns include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

A.2 Heterogeneity in reactions to *Leadership*

In this section, we briefly investigate heterogeneous reactions to treatments (see Tables A.3 and A.4). We do not find strong interactions of our *Leadership* condition and observable team characteristics such as group size, experience, the median age, share of males, or whether someone in the team took the walkie-talkie before ETR staff asked the team to do so. However, the interaction of speaking German and our leadership treatment turns out to be negative and statistically significant at the five-percent level for the probability to solve the task within 60 minutes (even though jointly, the coefficients *Leadership*, German, and the interaction are positive), but is statistically insignificant for the intensive margin ($p = 0.21$).

One particularly interesting aspect is whether teams in corporate bookings react differently to the treatment than teams in private bookings. On the one hand, teams of colleagues in corporate bookings (henceforth "corporate teams"), may be more likely to experience the endogenous emergence of a leader, as they may be used to a hierarchical organization through their work environment or may be more aware of the importance of leadership. On the other hand, one could argue that hierarchical structures are longer-lasting and well-defined among family and friends, therefore giving rise to more endogenous leadership formation among the latter. To illustrate potential differences between these groups further, we present separate cumulative distributions of finishing times in Appendix Figure A.1 in addition to the regression results shown in Appendix Tables A.3 and A.4, Column (4). It becomes clear that both private and corporate teams benefit from *Leadership*. Differences in treatment effects across these groups appear minor and turn out to be statistically insignificant (see Appendix Tables A.3 and A.4, Column (4)).

Figure A.1: CDFs of finishing time



Notes: The left panel shows the cumulative distribution of finishing times for private teams we asked to decide on a leader (*Leadership*) and without any intervention (*Control*). The right panel shows the same for corporate teams.

Table A.3: Team performance (completion, interactions)

	Completed within 60 minutes							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Leadership	0.112**	0.182	0.100	0.233	-0.032	-0.068	0.447***	0.201***
	(0.048)	(0.206)	(0.093)	(0.140)	(0.095)	(0.224)	(0.141)	(0.068)
Group Size	0.082***	0.092***	0.083***	0.084***	0.085***	0.083***	0.084***	0.082***
	(0.027)	(0.033)	(0.027)	(0.028)	(0.028)	(0.027)	(0.027)	(0.027)
Experience	0.142**	0.141**	0.130	0.143**	0.145**	0.146**	0.149**	0.142**
	(0.062)	(0.060)	(0.100)	(0.063)	(0.063)	(0.062)	(0.061)	(0.062)
Private	0.050	0.049	0.051	0.164	0.043	0.053	0.025	0.047
	(0.061)	(0.062)	(0.060)	(0.155)	(0.061)	(0.060)	(0.059)	(0.061)
Men Share	0.037	0.035	0.037	0.037	-0.149	0.032	0.032	0.027
	(0.091)	(0.091)	(0.091)	(0.091)	(0.149)	(0.089)	(0.093)	(0.093)
Median Age	-0.003	-0.003	-0.003	-0.003	-0.003	-0.006	-0.003	-0.003
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)
German	0.041	0.043	0.041	0.010	0.032	0.044	0.257**	0.046
	(0.106)	(0.106)	(0.106)	(0.130)	(0.105)	(0.107)	(0.117)	(0.108)
Walkie Talkie	-0.005	-0.005	-0.005	-0.009	0.005	-0.010	0.006	0.068
	(0.051)	(0.051)	(0.052)	(0.052)	(0.055)	(0.050)	(0.053)	(0.087)
Leadership x ...								
... Group Size		-0.016						
		(0.046)						
... Experience			0.018					
			(0.112)					
... Private				-0.152				
				(0.160)				
... Men Share					0.270			
					(0.163)			
... Median Age						0.006		
						(0.007)		
... German							-0.385**	
							(0.154)	
... Walkie Talkie								-0.117
								(0.093)
Mean in Control	0.442	0.442	0.442	0.442	0.442	0.442	0.442	0.442
Observations	281	281	281	281	281	281	281	281
Team Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Staff FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weekday and Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

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

Table A.4: Team performance (finishing times, interactions)

	Tobit: Finishing time							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Leadership	-2.551**	-6.909	-3.349	-3.887*	-2.316	1.933	-6.089**	-2.759
	(1.253)	(5.289)	(2.089)	(2.221)	(1.826)	(3.983)	(2.885)	(2.050)
Group Size	-1.907***	-2.549***	-1.886***	-1.919***	-1.911***	-1.920***	-1.908***	-1.907***
	(0.562)	(0.970)	(0.551)	(0.558)	(0.563)	(0.557)	(0.550)	(0.562)
Experience	-3.491**	-3.399**	-4.283**	-3.482**	-3.492**	-3.530**	-3.552**	-3.488**
	(1.425)	(1.423)	(2.087)	(1.445)	(1.427)	(1.402)	(1.436)	(1.430)
Private	-1.819	-1.758	-1.765	-3.127	-1.816	-1.935	-1.561	-1.815
	(1.350)	(1.353)	(1.331)	(2.522)	(1.357)	(1.340)	(1.403)	(1.356)
Men Share	-1.562	-1.467	-1.583	-1.555	-1.239	-1.557	-1.521	-1.535
	(1.375)	(1.394)	(1.396)	(1.370)	(2.792)	(1.380)	(1.375)	(1.398)
Median Age	0.094	0.092	0.096	0.092	0.094	0.190*	0.096	0.093
	(0.081)	(0.081)	(0.081)	(0.082)	(0.081)	(0.097)	(0.081)	(0.081)
German	-2.416	-2.431	-2.440	-2.069	-2.393	-2.448	-4.893**	-2.429
	(1.573)	(1.565)	(1.588)	(1.806)	(1.617)	(1.618)	(2.386)	(1.595)
Walkie Talkie	-0.148	-0.107	-0.144	-0.112	-0.168	-0.011	-0.251	-0.329
	(1.186)	(1.202)	(1.184)	(1.199)	(1.225)	(1.173)	(1.201)	(2.115)
Leadership x ...								
... Group Size		0.950						
		(1.199)						
... Experience			1.070					
			(2.530)					
... Private				1.688				
				(2.612)				
... Men Share					-0.447			
					(3.011)			
... Median Age						-0.142		
						(0.118)		
... German							3.998	
							(3.166)	
... Walkie Talkie								0.277
								(2.208)
Mean in Control	56.814	56.814	56.814	56.814	56.814	56.814	56.814	56.814
Observations	281	281	281	281	281	281	281	281
Team Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Staff FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weekday and Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table displays coefficients from Tobit regressions of finishing times on our treatment indicator (with Control as base category). All columns include room fixed effects. Each column indicates whether team controls (group size, share of male participants, experience with escape games, median age, language spoken, private versus team-building events, actively taken walkie-talkie), staff, weekday and week fixed effects are included. Standard errors in parentheses are clustered at the daily level, with significance levels * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

A.3 Team characteristics and choosing a leader

Table A.5, Column (1) shows whether team characteristics and the *Leadership* treatment affect the probability to select a leader before working on the task. In Column (2), we estimate the same model separately for each leadership sub-treatment (Motivation and Coordination). Column (3) estimates whether observable team characteristics predict the gender of the chosen leader. We find a (mechanical) negative relationship between the share of males and choosing a female leader as well as a positive relationship between median age and female leadership. Further, we find some indication that German speaking teams are less likely to choose a female leader. The latter result should however be taken with a grain of salt, as only a small minority of teams does not speak German.

Table A.5: Choosing a leader immediately

	Chose leader immediately (1)	Chose leader immediately (2)	Chose female leader (3)
Leadership	0.556***		
	(0.038)		
Motivation		0.562***	
		(0.051)	
Coordination		0.552***	
		(0.043)	
Group Size	-0.009	-0.008	0.001
	(0.035)	(0.035)	(0.076)
Experience	0.007	0.007	0.077
	(0.059)	(0.060)	(0.107)
Private	0.002	0.003	0.006
	(0.060)	(0.060)	(0.112)
Men Share	-0.107	-0.107	-0.917***
	(0.088)	(0.087)	(0.127)
Median Age	-0.003	-0.003	0.011**
	(0.004)	(0.004)	(0.005)
German	0.085	0.083	-0.556***
	(0.114)	(0.116)	(0.128)
Walkie Talkie	0.009	0.010	-0.040
	(0.045)	(0.045)	(0.091)
Mean in Control	0.000	0.000	-
Observations	281	281	81
Team Controls	Yes	Yes	Yes
Staff FE	Yes	Yes	Yes
Weekday and Week FE	Yes	Yes	Yes

Notes: The table displays coefficients from OLS regressions of whether a team chose a leader immediately (before they start working on the task) on our treatment (column (1): Leadership pooled, columns (2): Motivation and Coordination) indicator (with Control as base category) and OLS regressions of whether a team chose a female leader on team controls. All columns include room fixed effects. Each column indicates whether team controls (group size, share of male participants, experience with escape games, median age, language spoken, private versus team-building events, actively taken walkie-talkie), staff, weekday and week fixed effects are included. Standard errors in parentheses are clustered at the daily level, with significance levels * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

A.4 Results from customer survey

To analyze how teams perceived their experience and performance, Table A.6 presents the results from OLS regressions, as well as the 2nd stage from 2SLS regressions following the approach recommended in Angrist and Pischke (2008, p.142).[16] Each column uses a different survey question as the dependent variable, and these variables have been standardized to have mean zero and a standard deviation of one. Panel A reveals that the *Leadership* encouragement significantly affects perceived effort provision, motivation, and coordination. Panel B reveals even stronger results for choosing a leader on perceived effort provision, motivation and coordination.

Table A.6: Customer survey

	Price (1)	Satisfaction (2)	Effort (3)	Motivation (4)	Coordination (5)
Panel A. OLS (ITT)					
Leadership	0.016	0.020	0.455***	0.559***	0.325*
	(0.211)	(0.190)	(0.114)	(0.168)	(0.191)
Panel B. 2SLS (2nd Stage)					
Chose leader immediately	0.033	-0.102	0.543***	0.731***	0.454**
	(0.254)	(0.235)	(0.180)	(0.225)	(0.207)
Observations	135	135	135	135	135
Team Controls	Yes	Yes	Yes	Yes	Yes
Staff FE	Yes	Yes	Yes	Yes	Yes
Weekday and Week FE	Yes	Yes	Yes	Yes	Yes
Kleibergen-Paap Wald F (B)	98.75	98.75	98.75	98.75	98.75

Notes: The table displays coefficients from OLS (panel A) and 2SLS (panel B) regressions of answers in the customer survey on our treatment indicator (with Control as base category). All variables are standardized with mean zero and standard deviation of one. For 2SLS (Panel B) we follow the procedure outlined by Angrist and Pischke (2008): in a first step, we predict the probability of immediately choosing a leader using all control variables and fixed effects, as well as our treatment indicator in a Probit model. Then, we use these nonlinear fitted values as instruments in the second stage. All columns include room fixed effects. Each column indicates whether team controls (group size, share of male participants, experience with escape games, median age, language spoken, private versus team-building events, actively taken walkie-talkie), staff, weekday and week fixed effects are included. Standard errors in parentheses are clustered at the daily level, with significance levels * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

[16] Because filling in the customer survey was voluntary, we only include teams with complete responses.