The Importance of Team Governance Structures for Performance and Satisfaction

Dominik Michael Grothe

Dissertation

Ludwig-Maximilians-Universität Munich
Department of Economics
March 2023

The Importance of Team Governance Structures for Performance and Satisfaction

Inaugural-Dissertation
zur Erlangung des Grades Doctor oeconomiae publicae (Dr. oec. publ.)
an der Ludwig-Maximilians-Universität München

2023

vorgelegt von:
Dominik Michael Grothe

Referent: Prof. Dr. Florian Englmaier
Korreferentin: Prof. Dr. Lea Heursen
Promotionsabschlussberatung: 12. Juli 2023

Datum der mündlichen Prüfung: 05. Juli 2023
Berichterstatter: Prof. Dr. Florian Englmaier, Prof. Dr. Lea Heursen, Prof. Dr. Jana Gallus

ACKNOWLEDGMENTS

This journey would not have been possible without the support, help and encouragement of my advisors, colleagues, friends, and family. I am thankful to everyone who contributed to my research and to this wonderful time as a doctoral student.

First and foremost, I would like to thank my doctoral advisor, Florian Englmaier. His support, guidance and expertise helped me to grow as a researcher and person throughout the last couple of years. Furthermore, I would also like to thank my second advisor, Lea Heursen, from whom I have learned a lot about designing and conducting experiments. Finally, I am thankful to Jana Gallus, who was my host during my research visit at UCLA (University of California, Los Angeles) and has kindly accepted to serve on my doctoral committee.

I am indebted to the Elite Network of Bavaria, which provided generous funding for my doctoral position through IDK Evidence Based Economics.

This dissertation consists of four chapters, which are all part of a larger research agenda to answer several questions about team governance, team organization and team performance in non-routine analytical team tasks. Three of these chapters are the result of joint work with my highly valued co-authors Florian Englmaier (LMU Munich), David Schindler (Tilburg University), and Simeon Schudy (LMU Munich). I enjoyed working on these projects and benefited incredibly from the guidance and expertise of my co-authors. During this time, I had the opportunity to present our work at various conferences and workshops, which improved our projects and carried me forward as a person.

The different projects in this dissertation benefited a lot from the support of various research assistants. I would like to thank Tamina Straub, Oguzhan Bekar, Benjamin Schumann (Chapter 1), Simon Klein, Aloysius Widmann (Chapter 2), Silvia Castro (Chapter 1 and Chapter 2), Yutaka Makabe (Chapter 3 and Chapter 4), Sarah Birneder, Lisa Eitinger, and Martin Wiegand (Chapter 4) for their invaluable support. I am also thankful to Manuela Beckstein for her help and administrative support.

I would like to extend my thanks to a tremendous group of friends and colleagues, who contributed to the fact that I have greatly enjoyed my time as a doctoral student. This includes, most of all, Silvia Castro, Marvin Deversi, Carolin Formella, Svenja Friess, Lion Henrich, Michael Hofmann, Schanzah Khalid, and Anne Niedermaier. A very special word of thanks goes to my friend and former officemate Michael Kaiser. Without him, this time would not have been nearly as productive and beautiful as it was.

Finally, words cannot express my gratitude to my parents, Barbara and Michael, and my wife and love of my life, Magdalena. Thank you for your unconditional support and devotion.

Dominik Michael Grothe

München, 15. März 2023

Contents

List of Tables

List of Figures

PREFACE

> Employees are a company's greatest asset - they're your competitive advantage. You want to attract and retain the best; provide them with encouragement, stimulus and make them feel that they are an integral part of the company's mission.
>
> Anne M. Mulcahy

Over the last decades, the work environment has changed substantially. Moving away from routine, manual tasks that are performed individually, tasks have become more non-routine, complex, analytical and are frequently performed in teams (Autor et al., 2003; Autor and Price, 2013). Accordingly, teamwork is perceived as an essential requirement for success in modern firms (Bandiera et al., 2013; Weidmann and Deming, 2021).[1] This development, together with the general demographic trend and skills shortage, has led to new challenges for firms and their managers: They have to form teams, choose governance structures, and assign tasks such that team members become more productive and are at least equally or even more satisfied with their jobs at the same time (e.g. to reduce turnover). This task is particularly challenging for two reasons: First, causal evidence on the role of team composition and team governance structures for non-routine tasks is still scarce[2], and second, practitioners are not immune to biases, even if they could acquire relevant information (e.g. Kübler et al., 2018).

[1] The National Association of Colleges and Employers Survey highlights that firms require new employees to be able to work in teams (NACE, 2022).

[2] Recent exceptions are Hoogendoorn et al. (2013), who show positive effects of gender diversity on team performance in startup teams, and Englmaier et al. (2018), who show positive effects of bonus incentives on team performance in escape challenges.

The overarching theme of my dissertation is to provide causal evidence for the effectiveness of crucial determinants of team composition and team governance structures for team performance and satisfaction. I also investigate how these factors are perceived and potentially misperceived by practitioners. Using evidence from four field experiments, this dissertation documents insights for practitioners of organizational design and the fields of organizational and experimental economics.

According to the introductory quote[3], employees are the greatest asset of a company, and, particularly in times of demographic change and skills shortage, it is more important than ever to assess how teamwork can stimulate individuals' productivity while ensuring employee retention. To achieve this, it is crucial to understand and fulfill employee needs. There exist different approaches to apply Maslow's hierarchy of human needs to an organizational context.[4] In his paper "A Theory of Human Motivation", Abraham Maslow (1943) describes human motivation and needs in a hierarchy of five levels. The bottom level are physiological needs, which are followed by safety needs (second level), love and belonging (third level), esteem (fourth level), and finally self-actualization (fifth level). Before needs in higher levels can act as motivators, the needs in lower levels must be satisfied. Applying this scheme to the workplace, first of all, employees need a reasonable and regular income to fulfill their physiological needs. Once these basic needs are fulfilled, employees are motivated by the need of feeling safe (e.g. job security). Feeling like one belongs to the workplace and fits into a team is the next level in the hierarchy. Recognition and appreciation for the work an employee has accomplished represents the fourth level. Finally, when all other needs are fulfilled, employees are motivated by contributing to a higher goal and being challenged according to their skills and abilities without being overwhelmed. In general, team members may be located on different levels in the hierarchy and hence motivated by different needs. Accordingly, there is no one-size-fits-all approach to improve performance and satisfaction at the same time. Nonetheless, it is crucial to analyze the impact of team-level interventions on average team performance and satisfaction.

[3]The introductory quote's author, Anne M. Mulcahy (born October 21, 1952), served on various boards of directors and was CEO of Xerox Corporation. Among others, she was recognized as a great leader and powerful woman by the Chief Executive Magazine, Forbes Magazine, and the Wall Street Journal.

[4]Ruchi Kulhari, for example, matches career stages with stages of Maslow's hierarchy of needs (see https://www.forbes.com/sites/forbeshumanresourcescouncil/2021/06/10/maslows-hierarchy-of-needs-in-your-organization-how-to-support-your-employees-at-every-stage/?sh=748db7123b59)

There are two ways to identify the impact of an intervention, namely experimental and quasi-experimental methods. The main difference between these two approaches is the selection of a proper counterfactual. To select the counterfactual experimental methods randomize treatment assignment, while quasi-experimental methods make use of naturally occurring circumstances with sufficiently random treatment assignment.[5] Most common quasi-experimental approaches include difference-in-differences estimations (e.g. Card and Krueger, 1994), regression discontinuity designs (e.g. Thistlethwaite and Campbell, 1960), instrumental variables (e.g. Angrist, 1990) and matching estimators (e.g. Imbens et al., 2001). Experimental methods can be broadly separated into laboratory and field experiments. According to Harrison and List (2004), field experiments can further be divided into three different categories: Artefactual field experiments, framed field experiments and natural field experiments. While artefactual field experiments use an abstract framing and certain rules like in conventional lab experiments, they use a nonstandard subject pool in contrast to a standard subject pool of students. A framed field experiment exhibits the same characteristics as an artefactual field experiment, but either the commodity, task, or information set are of field context. Finally, a natural field experiment is characterized by a situation, in which subjects work naturally on these tasks and do not know that they are part of an experiment.

Using evidence from four field experiments, this dissertation analyzes the impact of three team-level interventions on performance and satisfaction as well as practitioners' perceptions regarding important determinants for team success. These studies are based on different experimental approaches, depending on the respective research question and the feasibility of using other approaches. The results of Chapter 1 (*The Efficacy of Tournaments*) and Chapter 2 (*The Value of Leadership*) are based on data from two natural field experiments, in which my co-authors and I use the unique opportunity to observe team performance in a setting, where participants naturally work on a non-routine analytical team task and are not informed that they are participating in an experiment.[6] The results of Chapter 3 (*(Mis-)Perceived Determinants of Team Success*) are based on an artefactual

[5] In 2021, David Card, Joshua D. Angrist and Guido W. Imbens received The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel for their contributions on how natural experiments can help to answer important societal questions (see https://www.nobelprize.org/prizes/economic-sciences/2021/popular-information/).

[6] For these studies, we cooperated with *Exit The Room*, which is a provider of real-life escape challenges. Escape challenges require cognitive, analytical and social skills and hence mirror defining features of many modern jobs (Deming and Kahn, 2018).

field experiment, in which my co-authors and I elicited practitioners' perceived importance of different attributes of team composition and team governance structures for team performance. Due to the lack of observational data on choices related to team composition and team governance structures, we use a discrete choice experiment to elicit the beliefs of Human Resources (HR) experts.[7] Chapter 4 (*Task Assignment*) is based on a framed field experiment, in which participants work on one of two different tasks. The study relies on a setting, in which a large number of people is working on two different tasks, where one could exogenously and randomly assign people to tasks. For this reason, I explicitly invite students for a one-time job that involves the preparation and analysis of data (i.e. regular tasks of research assistants).

The sequence of chapters in this dissertation follows the idea of first establishing causal evidence for the effectiveness of tournament incentives (Chapter 1) and the value of leadership (Chapter 2) for team performance in non-routine analytical team tasks. Thereafter, this thesis presents practitioners' perceptions and misperceptions of important determinants of team success (Chapter 3). Finally, focussing on a frequent task of leaders, this dissertation analyzes the impact of task assignment and task comparisons on performance and satisfaction as well as the efficacy of different task assignment procedures (Chapter 4). Taken together, this dissertation provides insights for practitioners and scholars based on empirical evidence from four different field experiments, which are described in more detail below.

Chapter 1[8] is based on evidence from a large-scale natural field experiment, where 1,728 participants in 378 teams work on a non-routine analytical task. While tournament incentives have ever been used to foster performance in innovation contexts, causal evidence for the efficacy of three important components innate to tournaments, namely salient team identity, social image concerns and monetary prizes, is scarce. Thus, my co-authors and I exogenously vary the existence of these components in a step-wise manner and evaluate their impact on team performance.[9]

[7] Discrete choice experiments allow to measure revealed preferences by exogenously varying choice attributes and letting participants choose between two (or more) hypothetical alternatives in a series of comparisons.

[8] A version of this chapter is accepted for publication in the Journal of Labor Economics (`https://doi.org/10.1086/725553`, accepted on April 19, 2023).

[9] Lazear and Rosen (1981) argue that tournaments, since they only require information on relative ranks instead of absolute performances, are a particularly attractive incentive scheme as they can establish efficient outcomes at lower costs.

We find that salient team identity alone does not lead to improved performances. Adding social image concerns improves performance at the top of the performance distribution, while further adding monetary incentives improves performance along the whole performance distribution. Furthermore, none of the components negatively affects the willingness to perform similar tasks again. Thus, this chapter shows that tournament incentives can substantially improve team performance in non-routine analytical tasks without harming overall satisfaction with the task.

Chapter 2[10] is based on evidence from a large-scale natural field experiment, where 1,273 participants in 281 teams work on the same non-routine analytical team task as in Chapter 1. While there is broad consensus regarding the importance of leadership (Antonakis et al., 2022), causal evidence of the actual value of leadership for teams performing a non-routine analytical task is scarce. This is mainly due to the fact that the presence of leadership is typically determined endogenously. To analyze the value of leadership, my co-authors and I exogenously vary the presence of leadership by encouraging randomly selected teams to choose a leader before they start working on the task.

We find that encouraging teams to choose a leader increases the probability of completing the task within the given time limit and reduces the time needed to complete the task. Different framings of leadership functions (motivation or coordination) affect team performance and team organization in a similar way. We further show that choosing a leader is indeed improving performance.[11] The choice of a leader does not reduce satisfaction with the task and likely improves coordination among team members (i.e. decentralized information acquisition and problem solving).

Chapter 1 and Chapter 2, together with the findings in Englmaier et al. (2018), enabled my co-authors and me to exploit the unique opportunity to analyze the perceived importance of team composition and team governance structures for team success and contrast it with actual performance data of 1,062 teams. The results are presented in Chapter 3, which is based on evidence from a large-scale discrete choice experiment. In this study, 3,000 HR experts and 3,000 people from a general population sample had to predict which of two teams, with different team compositions and governance structures, performs better in two different non-routine analytical tasks.

[10] Parts of this chapter will be included in a revision of Englmaier et al. (2018) for peer-review at the Journal of Political Economy (not submitted at the time of writing).

[11] Using an instrumental variables approach following Angrist and Pischke (2008).

We find that HR experts hold qualitatively accurate beliefs, but substantially underestimate the value of leadership. Partially depending on the gender of the expert, we identify implicit biases against leadership, in particular female leadership. Regarding the efficacy of (female) leadership, second-order beliefs tend to be more optimistic and gender-specific biases are less pronounced. Additionally, we show that the general population sample holds similar beliefs about the importance of team composition and team governance structures for team success in non-routine analytical tasks.

Chapter 4 presents a framed field experiment to investigate the impact of task assignment among co-workers on performance and satisfaction. Allocating tasks to team members is a frequent exercise for leaders and an important dimension for work performance, work motivation and satisfaction with the work. While there is evidence that workers compare their own task with the task of close co-workers, causal evidence for the impact of task comparisons and the effectiveness of different task assignment procedures is scarce. In the run-up to the experiment, I conducted a survey with 400 practitioners to elicit their beliefs regarding the impact of task assignment and task comparisons as well as different task assignment procedures on performance and satisfaction. The results illustrate the heterogenous beliefs of practitioners, which reinforces the need for an empirical investigation of this relationship.

To ensure efficient use of research funds and using tasks of field context, the start of this experiment relies on the completion of an ongoing RCT.[12] There, my co-authors and I analyze microaspects of leadership by exogenously varying the existence of an endogenously or exogenously chosen leader. We collect performance, audio and tracking data as well as detailed data on participants' characteristics and a measure of their creativity. The analysis of the audio data requires two distinct tasks: First, the audio data have to be transcribed, and second, the transcribed data have to be evaluated. This serves as an ideal opportunity to analyze the impact of task assignment and different task assignment procedures on performance and satisfaction. Since these are regular tasks for research assistants, students are the relevant population and explicitly invited to perform them as part of a one-time job. In the experiment, groups of two close co-workers are randomly allocated to either work on the same or different tasks. In the latter case, the task assignment procedure is exogenously varied: Tasks are either randomly assigned, allocated

[12]Data collection is expected to be completed in April 2023. After that, I can start with the roll-out of this experiment. This chapter presents the motivation, experimental design, planned data collection and analyses.

according to (self-evaluated) perceived productivity or preferences, or are self-assigned among co-workers. Based on predictions from a stylized agency model, I expect positive effects for advantaged (i.e. working on the preferred task or the task for which they expect to perform better) and negative effects for disadvantaged workers (i.e. working on the less preferred task or the task for which they expect to perform worse), when assigned to a different task than the co-worker instead of working on the same task. Furthermore, I expect self-assignment, as long as co-workers reach an agreement, to be the most effective assignment procedure. The efficacy of the other procedures depends on whether they lead to an efficient allocation and whether they are accepted as reasonable justifications.

Taken together, this dissertation delivers important insights for scholars and practitioners alike. Chapter 1 and Chapter 2 establish causal evidence for the efficacy of tournament incentives and the value of leadership for performance in non-routine analytical tasks with a clearly specified goal and deadline. Accordingly, tournament incentives and the encouragement to endogenously choose a leader can serve as cost-effective tools for improving team performance without making teams less satisfied with the task. Using the evidence from Chapter 1 and Chapter 2, Chapter 3 reveals perceived and misperceived determinants of team success in non-routine analytical team tasks. The results indicate a substantial underestimation of the value of leadership, particularly female leadership. Making HR experts aware of their biases and reducing their misperceptions, particularly with respect to the value of (female) leadership, offers an opportunity to make team performance even more successful. Finally, Chapter 4 seeks to inform about the impact of task assignment and task comparisons on performance and satisfaction. This study further examines the importance of perceived productivity and preferences for perceptions of task differences and analyzes how different task assignment procedures might mitigate these effects.

Chapter 1

The Efficacy of Tournaments for Team Performance in Non-Routine Analytical Team Tasks [1]

Abstract

Tournaments are often used to improve performance in innovation contexts. Tournaments provide monetary incentives but also render teams' identity and social image concerns salient. We study the effects of tournaments on team performance in a non-routine task and identify the importance of these behavioral aspects. In a natural field experiment ($n > 1,700$ participants), we vary the salience of team identity, social image concerns, and whether teams face monetary incentives. Increased salience of team identity does not improve performance. Social image motivates mainly the top performers. Additional monetary incentives improve all teams' outcomes without crowding out teams' willingness to explore or perform similar tasks again.

[1] This chapter is based on joint work with Florian Englmaier (LMU Munich), Stefan Grimm (LMU Munich), David Schindler (Tilburg University) and Simeon Schudy (LMU Munich). A version of this chapter is accepted for publication in the Journal of Labor Economics (https://doi.org/10.1086/725553, accepted on April 19, 2023).

1.1 Introduction

Ever since the seminal contribution of Lazear and Rosen (1981), there has been great interest in tournaments to foster performance and innovation (cf. the overview in Lazear and Oyer, 2012).[2] Lazear and Rosen's original argument for the attractiveness of tournaments relied on the fact that tournaments can establish efficient outcomes at lower costs, since tournaments only require information on relative ranks instead of absolute performance. However, in innovation contexts, in which teams derive status from developing innovative solutions, tournaments include additional and important behavioral features, rendering them attractive for improving performance. First, tournaments naturally increase the salience of team identity because teams are explicitly identified (e.g., by a ranking of teams, departments, brand, or company names). Second, as the rankings are observable, tournaments may substantially intensify status-related image concerns. Prior research in psychology and economics has documented that both identity (see, e.g., Tajfel and Turner, 2001; Akerlof and Kranton, 2000; Chen and Chen, 2011) and image concerns (see, e.g., Kluger and DeNisi, 1996; Kosfeld and Neckermann, 2011; Fershtman et al., 2006; Ball et al., 2001; Moldovanu et al., 2007; Bursztyn and Jensen, 2017) can play a crucial role in human behavior. However, much less is known about their role in the efficacy of tournaments in complex, non-routine analytical tasks, which have become ubiquitous in modern economies and characterize many work environments in innovation contexts (see, e.g., Autor et al., 2003; Autor and Price, 2013). Since understanding the relative importance of these aspects enables a cost-effective design of incentives, the aim of the present study is twofold: to investigate the efficacy of tournaments with prizes in non-routine team tasks, and to determine the importance of behavioral aspects vis-a-vis monetary rewards.

This study exploits a unique field setting to understand the importance of salient identity, image concerns, and prizes in tournaments involving complex teamwork. We conduct a large-scale field experiment to identify the causal effects of these components on team performance in real-life escape room challenges, in which teams have to solve a series of cognitively demanding tasks in order to succeed. These tasks are popular world-

[2]Early examples of innovation competitions were the "longitude rewards", a system of inducement prizes offered by the Government of Great Britain for a practical and straightforward method to precisely determine a ship's longitude at sea. These rewards were granted by Parliament in 1714 and were administered by the newly created Board of Longitude. Brunt et al. (2012) and Khan (2015) provide more details on the role of inducement prizes in innovation.

wide both among private teams seeking a complex team challenge and companies which use them for team building and recruiting purposes.[3] Escape challenges require cognitive skills, analytical and critical thinking, as well as social skills such as communication and collaboration. Thereby, they mirror defining features of many modern jobs (Deming and Kahn, 2018). Teams face a series of complex problems, need to collect and recombine information, and think outside the box. The tasks are interactive, as team members have to collaborate with each other, discuss possible actions, jointly develop ideas, and test their hypotheses. Hence, escape challenges encompass important elements of production in the "ideas sector" of the economy (see, e.g., Autor et al., 2003; Autor and Price, 2013) and require abilities, which modern employers consider of utmost importance (Deming, 2017; Casner-Lotto and Barrington, 2006; Jerald, 2009). Additionally, escape challenges allow for an objective measurement of team performance (the time spent until completion). At the same time, the team challenge provides space for team identity and image concerns to matter, as teams often proudly document their participation on a local "wall of fame" on site. Finally, the setting allows for exogenously manipulating important tournament characteristics such as the salience of identity, image, and instrumental concerns across a large number of teams.[4]

To identify the importance of team identity, image concerns, and monetary prizes, we randomly allocated participating teams to one of four conditions, which introduced these features in steps (such that each additional step also comprised the treatment components of the previous step). To analyze the importance of salient team identity, we first compare a no intervention condition *Control*, in which teams have no team name, with a condition, in which we ensure salient team identity by asking teams to explicitly discuss, and jointly choose, a team name they identify with. Since in most business contexts, teams already have some team (or brand) name they identify with, this condition also serves as a meaningful comparison group for the investigation of the additional effects of image concerns due to public rankings and instrumental concerns due to monetary prizes. Our second treatment condition focuses on image concerns and introduces a public ranking for all teams (using self-chosen team names) and our third treatment

[3] Escape challenges are also used for education purposes of IT and Engineering students (see, e.g., Borrego et al., 2017) and prior research has used other unique opportunities to study competition in tournaments, e.g., data from sports (see, e.g., Brown, 2011; Brown and Minor, 2014).

[4] A more extensive discussion of the features of the setting and the task and the responsiveness of team performance to bonus incentives is provided in Englmaier et al. (2018).

condition is a classical tournament, in which teams are publicly ranked (with self-chosen team names) and the best team receives a monetary prize.

We find that introducing salient team identity alone is not sufficient to improve team performance, but adding image concerns in the form of rankings appears to matter. When a treatment features a public ranking, teams tend, on average, to solve the task more quickly, which is mostly driven by the top performers. Those below the top quantiles are, however, similarly likely to complete the task compared to teams whose performance is not publicly ranked. Introducing a monetary prize in addition to the public ranking substantially increases the likelihood of succeeding within the given time limit. Prizes boost performance at the top but also along the lower quantiles of the performance spectrum. Overall, the tournament with a monetary prize and public ranking (using self-chosen team names) increases completion rates by more than 20 percent (almost 12 percentage points) as compared to *Control*, and reduced finishing times by more than 3 minutes (remaining times are almost doubled).

These findings contribute to the recent field work on tournaments, incentives, and teamwork in non-routine analytical tasks and complement findings from laboratory experiments on "closed-form" creative tasks.[5] First and foremost, we provide novel field-experimental evidence on the causal effects of three major components innate to tournament incentives (salience of team identity, image concerns, and instrumental concerns) for performance in non-routine, analytical team tasks. In this way, we systematically advance earlier field work that studied rank versus monetary incentives in *routine* tasks. Findings in the context of routine tasks indicate that tournaments with and without prizes can affect team performance, particularly when team identity is present. For instance, Delfgaauw et al. (2013) compare rank and monetary incentives in retail chains and document that sales competitions have a positive effect on sales growth, but only in stores where the store's manager and a sufficiently large fraction of the employees have the same gender (a proxy for stronger team identity). Our setting allows us to implement a treatment condition that exogenously assures salient team identity, and sheds light on how image concerns due to rankings, and instrumental concerns due to prizes affect performance in a non-routine task. Our results show that it is indeed the introduction of

[5] "Closed-form" creative tasks in the context of business innovation are for example characterized by specific goals such as enhancing a technological process, reducing costs, or refining an existing product. For a detailed discussion of open versus closed-form creativity see also Charness and Grieco (2019).

competition that fosters performance while assuring salient identity alone was ineffective.

In terms of public rankings and prizes, we also complement work by Bandiera et al. (2013) which focused on the productivity of fruit-pickers. In their setting, team rankings led to stark selection into teams based on team members' performance potential (rather than friendship networks) and reduced performance, due to an increase in free-riding. Tournaments with prizes had similar effects in terms of selection, but yielded additional effort provision within teams, which offset the negative effects of free-riding. Our study is novel and different to previous work in several important ways: First, we focus on a non-routine team task and vary incentives across the existing teams, excluding selection into teams based on incentives by design. Second, our setting allows us to vary the salience of team identity in a natural way without introducing competition. Third, while in previous work rankings are often informative about income differences (e.g., when teams are paid based on a piece-rate), our study isolates non-instrumental image concerns when introducing the public ranking. Excluding selection based on incentives and instrumental concerns, we find that introducing rank incentives has positive effects on performance. In contrast to studies on performance rankings in repeated settings (Blanes i Vidal and Nossol, 2011; Barankay, 2012; Ashraf et al., 2014; Bursztyn and Jensen, 2015; Delfgaauw et al., 2020; Ashraf, 2019; Blader et al., 2020), which sometimes document discouraging effects of relative performance rankings, we focus on the pure effect of the introduction of tournament incentives. Doing so, we show that the mere existence of tournament incentives (with and without prize) does not curb the preference for performing similar tasks again.

Studying non-routine tasks, we also complement recent laboratory studies focusing on the causal effects of incentives in creative tasks. Incentives have been discussed as potentially crowding out intrinsic motivation (e.g. Deci et al., 1999; Eckartz et al., 2012; Gerhart and Fang, 2015; Hennessey and Amabile, 2010). However, recent evidence suggests a more differentiated picture. In a laboratory experiment, Laske and Schroeder (2016) analyze incentives for the creativity of individuals, which they measure along three dimensions: quantity, quality, and originality of ideas. They compare piece-rate incentives for quantity alone, quantity combined with quality, and quantity in combination with originality, and a fixed wage condition. In their setting, incentives significantly affect the quantity and average quality of ideas, but not the average originality. Morgan et al. (2020)

find that performance-based incentives increase team effort in Fermi problems (Ärlebäck and Albarracín, 2019) but do not result in better guesstimations. Bradler et al. (2019) use a large-scale laboratory experiment to analyze the impact of tournament incentives and wage gifts on creativity. While tournaments substantially increase creative output, with no evidence for crowding out of intrinsic motivation, wage gifts are ineffective. Charness and Grieco (2019) analyze incentives for "open-" and "closed-form" creative tasks in the laboratory. Their results indicate that monetary incentives effectively stimulate creativity only in tasks with specific ex-ante goals ("closed-form") but not in creative, yet less well-defined tasks ("open-form"), whereas a ranking is effective in both types of tasks. In another laboratory experiment that arose simultaneously to our work, Charness and Grieco (2023) analyze the relationship between performance pay, corporate culture, and "closed-form" creativity. Akin to the escape challenge, they use tasks with specific ex-ante goals, and compare (among others) a treatment without performance incentives (flat pay), a group-ranking treatment without performance incentives (flat pay + ranking) and a group treatment with group-ranking and performance pay proportional to the team's rank (performance pay + ranking). Similar to our results, they observe positive effects of rankings and additional monetary rewards.

Finally, we also link to field work on creative production. Gross (2020) documents that increased competition can foster creative production of individual logo designers, but heavy competition drives designers to stop producing logos altogether. In a similar vein, Casas-Arce and Martinez-Jerez (2009) show that the introduction of sales contests fosters effort, while incentives weaken with an increase in competition (i.e., more participants). Complementing the above findings, our results provide important field-experimental evidence on the efficacy of incentives for non-routine analytical team tasks. Focusing on teamwork that requires the forming and testing of hypotheses to come up with the solution to a complex closed-form problem, we show that tournaments can stimulate performance in these goal-oriented tasks, both due to concerns for social image and instrumental concerns. We observe a robust performance-enhancing effect of rankings for the very top and of monetary prizes for all participating teams. At the same time, we do not observe negative side effects when offering these incentives. Teams neither request more external help to arrive at the solution nor do they request help earlier. In line with field evidence that focuses on incentives for idea creation (Gibbs et al., 2017), and laboratory evidence on "closed-form" creative tasks, the findings from our natural

field experiment suggest that incentives can foster performance in non-routine analytical team tasks with a specific goal. Lastly, we do not detect statistically significant effects on teams' revealed preferences for performing a similar task again: teams in conditions encompassing a ranking or a monetary prize are not less likely to purchase a voucher for future participation; if anything, our results point in the opposite direction.

The rest of this paper is structured as follows. Section 1.2 will describe the setting and our experimental design in more detail. Section 1.3 provides the results from the experiment, Section 1.4 discusses other possible mechanisms through which the three non-control conditions could affect performance, and Section 1.5 concludes.

1.2 Experimental design

1.2.1 The field setting

For this study, we collaborated with *ExitTheRoom* (ETR), a provider of real-life escape room challenges and conducted our natural field experiment (Harrison and List, 2004) at the facilities of ETR in Munich, Germany.[6] The location offers three differently themed rooms and teams face a time limit of 60 minutes.[7] Teams can see their remaining time on a large screen in their room and if a team manages to succeed within the time limit, they win. If time runs out before the team completes all quests, they lose. Teams participate in these challenges with the aim of succeeding before the deadline, and are proud of finishing the task quickly, which is also reflected by the fact that many participants write their finishing times on the walls of the entrance area of our collaboration partner. Further, as teams do not know how many quests the challenge consists of, teams naturally aim for succeeding quickly.[8] If teams get stuck, they can request up to five hints via a walkie-talkie. Hint-taking involves no explicit costs (neither monetary nor in terms of the remaining time). However, as the number of allowed requests for a hint is limited, there are opportunity costs of asking for assistance. ETR staff provides hints upon request but never gives the immediate solution to a (sub)task. Instead, they only include

[6] For more information, see their website at `https://www.exittheroom.de/munich`.

[7] 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 have to be found. *Zombie Apocalypse* requires teams to find the correct mix of liquids, an anti-zombie potion, before time runs out.

[8] Note that there is no entertainment value of simply waiting in the room without making any progress. In this setting, potential task utility merely stems from exploring the rooms and thereby making progress.

vague clues regarding the next required steps. At the very end, either after completing the task or reaching the time limit, ETR staff offers teams the opportunity to purchase a voucher for future participation at a reduced rate.

ETR provides a rich setting containing the key characteristics of modern non-routine analytical teamwork. Teams have to carry out a series of cognitively demanding tasks in which they need to acquire and combine information and develop and exchange ideas with their team members. Akin to environments in innovation contexts, teams are proud to succeed but the observability of co-workers' cognitive effort provision is limited (rendering the task prone to free-riding). Thus, the setting leaves room for team identity and image concerns to matter and constitutes an excellent environment for a natural field experiment.

Our setting reflects important characteristics of modern teamwork but also involves some caveats. First, teams solving the escape challenge choose to perform the task and likely derive task utility. While such selection is less common for traditional working environments, highly educated workers appear to deliberately self-select into occupations based on the interesting, non-routine nature of the tasks the occupation involves (Autor and Handel, 2013). Second, the effectiveness of tournament incentives may depend on a workers' motivation, which may not solely stem from the task itself, but also from salient greater goals that are missing in escape challenges. Importantly, Englmaier et al. (2018) show that monetary incentives are effective in the same escape setting, independent of differences in worker motivation and self-selection into the task. Finally, the escape challenge involves a complex problem with one final solution whereas complex problems in work environments may be multi-dimensional and in principle allow for several possible solutions. However, in innovation and business contexts, deadlines and budget-constraints often render one solution favorable, and the nature of the escape challenge mirrors this idea. It offers multiple ways to arrive at the (one) final solution and thereby allows us to study how tournament incentives motivate workers to produce the best possible solution within a given amount of time. As such, the escape challenge resembles the idea of closed-form creativity (Charness and Grieco, 2019), in which teams face a complex task with a well-delineated goal (as opposed to an open-form task that may not envision a specific final outcome). Thus, it can be reflective of modern work tasks in the context of business innovation, which may for example focus on the enhancement of

a technology process or the development of new ideas that solve a well-defined problem subject to time constraints.

1.2.2 Procedures and treatments

Our field experiment was conducted with 1,728 customers in 378 teams at *ExitTheRoom's* Munich location between April and July 2018 during their regular opening hours from Monday to Friday. Teams booked and paid online in advance. Upon arrival on-site, ETR staff welcomed the teams and delivered a standard introduction, laying out the story behind the specific room and explaining the task's rules.

To avoid contamination, we randomized treatment arms on a weekly level.[9] ETR staff implemented the different treatments after delivering the introduction. The choice of our experimental treatment variations was guided by the previous literature comparing tournaments with and without prizes (Barankay, 2012; Charness and Grieco, 2019), as well as by tournament designs in practice, which often involve rankings of team names that relate to teams' identity (e.g, the Netflix Prize). Hence, we focus on three components innate to tournaments: salient team identity (through team names), image concerns (due to being ranked), and instrumental concerns (due to prizes). Varying these three components independently would have resulted in a full factorial $2 \times 2 \times 2$ design with eight experimental conditions. However, our collaboration partner considered treatment variations in which we would i) publish team names without a ranking, ii) rank teams without a name teams can identify with, iii) or assign a prize to the best team without a public ranking with iv) or without team names (due to lack of transparency) incongruous. Thus, we opted for four experimental conditions which step-wise introduce team names, rankings, and prizes. These, we believe, cover also many applications relevant for practitioners, as prizes often involve public rankings, and public rankings usually require a unique and meaningful team identifier.

In our *Control* condition (112 teams), teams were not subject to any intervention and started working on the task directly after receiving the standard introduction. As tourna-

[9]ETR shared booking data from the first two weeks of our study period with us. This data reveals that more than 90% of the teams had already booked a slot in a given week before the first session in that week was conducted. Participating teams were not informed about the study and were thus unaware that we randomized at the weekly level as well as that there were different treatment arms. Learning about these aspects within the natural setting required repeated participation in at least two rooms in two different weeks, which disqualified the team's performance from our analyses. We identified six repeated (out of a total of 384) performances that are not included in our data.

ments render team identity salient by explicitly identifying them by their name, brand, or company, our first treatment condition, *T1 (Identity)* (85 teams), was designed to increase the salience of team identity in a natural way, without adding any competitive aspects. Following the idea in Ai et al. (2022), in which the company DiDi (a leading transportation platform) explicitly used the creation of team names by team members to increase team identification, we asked teams to jointly deliberate on a team name to be used for communication during the task with ETR staff via the walkie-talkie.[10] Teams were free to choose any name all members identified with, and were actively engaged in choosing the team name.[11]

To study the effects of introducing image concerns through competition, we implement our second treatment condition *T2 (Identity, Rank)*, 94 teams. Based on the idea that people care about being ranked per se (Charness et al., 2014; Charness and Grieco, 2019), and thus also about the rank of their team, *T2 (Identity, Rank)* includes a weekly tournament for teams facing the exact same challenge (i.e., the same of three rooms) without a prize.[12] In the same manner as in *T1 (Identity)*, we also asked team members in *T2 (Identity, Rank)* to select a team name. In addition, we informed teams that a ranking of the current week's teams would be publicly shown on ETR's Facebook account the following Monday (for an example, see Figure 1.A.1), where teams were ranked by room according to their finishing times with their team name. All teams that did not complete the task were assigned the same rank. Although the ranking did not reveal which team contained which members, team members were free to tell others about their team's performance, and some individuals indeed engaged with the weekly Facebook post using their real names (see also Figure 1.A.1).

Lastly, treatment *T3 (Identity, Rank, Prize)*, with 87 teams, exhibited the same features as *T2 (Identity, Rank)*, but in addition offered a prize of 150 Euro for the best team in a week (separately for each room). Winning teams were contacted by e-mail (simultaneously with the publication of the ranking) and invited to pick up the reward at the facilities of ETR at their earliest convenience. Incentives were large relative to the price paid for

[10] In *Control*, ETR staff referred to the team member with the walkie-talkie as "you".

[11] Thus, our treatment rendered the sense of belonging to a group salient instead of exogenously assigning an arbitrary team identity (see also the discussion in Sen, 2007; Chowdhury et al., 2016).

[12] As teams who booked the same room (usually several days in advance) do not encounter each other on site, and teams working in different rooms in overlapping time slots do not compete with each other, teams are unlikely to form informed priors about their potential competitors.

participation (which ranged between 99 and 129 Euro depending on the size of the team) and thus also salient.[13]

1.2.3 Outcome measures and sample characteristics

Our final sample consists of 373 teams (1,705 individuals, see Table 1.1).[14] We collected observable information related to team performance and background characteristics for all teams. These include time needed to complete the task, number and timing of requested hints, team size, gender and age composition of the team, team language (German or English), prior experience with escape rooms, and whether the customers came as a private group or were part of a corporate team-building event.[15] Further, we recorded the names of the teams in all treatments apart from *Control* (where teams did not choose a name).

Our primary outcome variable is team performance, which we measure by 1) whether teams completed the task within the time limit of 60 minutes, and 2) the time needed to complete the task. Exogenous variation in the salience of team identity, image concerns, and instrumental concerns allows us to estimate the causal effects on these outcomes. Furthermore, we analyze the impact on two secondary outcome variables: the willingness to explore original solutions (which we measure inversely by the number of hints a team has taken) and a team's interest to perform a similar task again (which we measure using the probability of purchasing a voucher for future participation at ETR at a reduced rate immediately after performing the task).

Table 1.1 provides an overview of team characteristics across treatments (team size, gender, age composition, team language: German or English, prior experience with escape rooms, and whether the team came as a private group or were part of a corporate team-building event). Accounting for multiple hypotheses testing following List et al.

[13] For the role of salience for incentives, see also Englmaier et al. (2017).

[14] During data collection, ETR's operation became inhibited after suffering from water damage resulting from a burst pipe in the building. The water damaged the electronics in the room *The Bomb*, leading to its use between June 18 and June 20 being reduced. In total, five teams in treatment *Prize* were affected before full functionality could be restored. To avoid capturing any effects on performance this may have had, we exclude these observations from the main analyses. We provide robustness checks showing that our results do not hinge on this decision in Table 1.A.9.

[15] To preserve the character of being a natural field experiment, we did not interfere with ETR's standard procedures. Therefore, we could not explicitly elicit the participants' ages. Instead, the age of each participant was estimated based on appearance to be either 1) below 18 years, 2) between 18 and 25 years, 3) between 26 and 35 years, 4) between 36 and 50 years, 5) 51 years or older. As we are interested in the behavior of adults (and in accordance with our IRB approval) we did not include teams with minors in our study.

Table 1.1: Sample size and characteristics

	Control - Mean (SD)	*T1* *Identity* Mean (SD)	*T2* *Identity, Rank* Mean (SD)	*T3* *Identity, Rank, Prize* Mean (SD)
Group Size	4.52 (1.01)	4.41 (0.95)	4.69 (1.01)	4.67 (1.01)
Experience	0.62 (0.49)	0.78 (0.42)	0.71 (0.45)	0.68 (0.47)
Private	0.79 (0.41)	0.89 (0.31)	0.85 (0.36)	0.89 (0.31)
Men Share	0.47 (0.28)	0.41 (0.28)	0.49 (0.30)	0.44 (0.30)
Median Age	32.88 (9.81)	30.26 (7.64)^{b}	33.69 (8.47)^{a}	31.47 (9.37)
German	0.89 (0.31)	0.99 (0.11)	0.94 (0.25)	0.96 (0.19)
Observations	112	85	94	82

Notes: Rows report means on the group level. Group size denotes the number of team members. Experience is a dummy for teams with at least one member who experienced an escape room challenge before. Private is a dummy whether a team participates as a private event (1) or whether the team belongs to a team building event (0). Men Share refers to the share of male team members. Median Age is defined as the median of all participants' guessed age categories' midpoint in a team. German is a dummy for German-speaking (1) or English-speaking (0) teams. Standard deviations in parentheses. Stars indicate significant differences to Control (p-values adjusted for multiple hypothesis testing following List et al. (2019), with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$); $\{a,b,c\}$ indicate differences to {*Identity*, *Rank*, *Prize*} at the ten percent level.

(2019), none of the observable characteristics differs significantly from *Control*. The only statistically significant difference (at the ten percent level) occurs for teams' median age (estimated by our RAs) when comparing *Identity* and *Rank*. We thus will show regression results with and without team characteristics as controls.

1.2.4 Hypotheses

The sense of identity and belonging is a fundamental human need (see, e.g., Baumeister and Leary, 2017). Experimental evidence from the laboratory suggests that salient team identity can alter cooperation and coordination within groups as well as reciprocity among agents, all of which are crucial for successful performance in the task at hand. For instance, Chen and Li (2009) use a (near) minimal group design and find that participants are 19 percent more likely to reward an in-group match for good behavior but 13 percent less likely to punish an in-group match for misbehavior. Drouvelis and Nosenzo (2013) provide evidence that group identity is beneficial in contexts that allow for leading by example, and Eckel and Grossman (2005) show that team identification may limit individual shirking and free-riding in environments with the character of a public good (in particular when paired with joint activities such as group problem-solving). Further, identity has been shown to affect group coordination and conflict (Chen and Chen, 2011; Chen et al., 2014; Chowdhury, 2021).

Our design focuses on the salience of team identity. While pre-existing groups arrive on the premises of our collaboration partner, jointly elaborating on and choosing a team name renders team identity salient. Our approach reflects current business strategies pursued by companies relying on structures based on agile teams rather than strict hierarchical structures.[16] In our context, we thus expect performance improvements when a team's identity is rendered more salient.

Hypothesis 1 *Rendering team identity more salient by asking team members to jointly deliberate on and choose a team name improves team performance.*

Competition between teams may reduce free-riding within each team as workers may care about their image, and change their behaviors based on how they are perceived by others. For instance, Tan and Bolle (2007) find that cooperation rates within teams (in laboratory public goods games) increase when outcomes are compared to other teams. Field-experimental evidence from individual routine tasks shows that non-instrumental rewards which encompass image value can substantially improve performance (Kosfeld and Neckermann, 2011). Further, Restivo and Van De Rijt (2012) show that informal rewards can raise contribution levels of high-performing individual contributors at Wikipedia. Studies on team performance in routine tasks suggest that rank incentives can substantially affect image concerns, and thereby team composition and performance. While changes driven by image concerns do not necessarily result in better performance (see, e.g., Bandiera et al., 2013; Kosfeld et al., 2017), positive effects have been observed in environments in which team identity was likely to be strong and salient (Delfgaauw et al., 2013). In line with these findings, we thus hypothesize that image concerns can boost performance (in addition to identity), also in non-routine tasks. Furthermore, prior research has documented that positive performance effects of symbolic rewards are particularly effective for top performers (Kosfeld and Neckermann, 2011). It thus seems reasonable to assume that image concerns may have different effects on teams depending on their relative likelihood of being ranked high. Teams that are expected to perform well based

[16] Based on insights from social and applied psychology (see e.g. Van Knippenberg, 2000; Van Dick et al., 2006) suggesting a strong positive relation between organizational identification and organizational citizenship, many firms emphasize team identity as an important factor for success and explicitly encourage the choice of a team name (see for example *Calabrio*, `https://web.archive.org/web/20210123010704/https://www.calabrio.com/wfo/workforce-management/boost-belonging-motivation-through-team-names/` and Ye et al., 2022).

on observable characteristics (e.g., because they are particularly able, more experienced, or particularly motivated to perform well) may show a stronger reaction to the public ranking than teams that are expected to perform worse (e.g., because they are less able, less experienced or less motivated to perform well). As lower ranks in our weekly competitions were likely to pool several teams failing to complete the task, teams at the bottom end of the performance distribution are likely to expect lower marginal image returns to effort. We thus hypothesize that positive performance effects of rank incentives are observed particularly for the upper quantile of the performance distribution.

Hypothesis 2 *Strengthening image concerns by implementing public rankings improves team performance, particularly for top performing teams.*

Field experiments randomly assigning teams to tournaments with monetary prizes or other instrumental rewards have so far mainly focused on routine tasks. For example, Erev et al. (1993) showed that tournament incentives can help teams of orange pickers to overcome problems of free-riding innate to environments that require voluntary contributions. Blimpo (2014) extends this positive link to learning outcomes and finds substantial and positive effects of tournaments with monetary prizes when teams of students compete across schools. Similarly, positive effects are also observed when tournaments involve non-monetary prizes (grade improvements) that have instrumental value (Bigoni et al., 2015). In line with expected image and instrumental returns from effort, such tournaments increase the performance of good students while they often appear less effective for students at the lower end of the performance distribution (De Paola et al., 2012). In the context of production, Delfgaauw et al. (2013) provide evidence from sales team competitions with and without prizes in discount stores. They observe positive effects of competition, both for tournaments with ranks only and tournaments with prizes. However, they find no evidence that financial rewards led to additional performance improvements, potentially due to strong image concerns and related ceiling effects or due to perceived instrumental values of ranks for employees (e.g., better perceived career opportunities or lower likelihood of job loss). Given the evidence discussed above, we expect that the introduction of prizes further improves team performance (as compared to tournaments without prizes).

Hypothesis 3 *Adding a monetary prize to the rank tournament improves team performance.*

The development of our hypotheses reflects the idea that salience of team identity, image concerns, and instrumental concerns are three major components innate to typical tournament incentives. We hypothesized that a public ranking introduces image concerns to a setting in which teams with salient team identity perform, and a monetary prize introduces instrumental concerns in settings in which teams otherwise compete for ranks. Alternatively, one could also hypothesize that image concerns through a public display of team names may interact with feelings of team identity and thus trigger an additional performance increase through stronger feelings of identity. Similarly, adding monetary prizes may additionally alter image concerns (or team identity). In other words, teams may perceive the value of appearing first in the public ranking differently, because monetary prizes may either crowd out parts of the image motivation or increase the image value of being first in the ranking. While we consider identity-strengthening aspects of additional image and instrumental concerns less likely in environments with otherwise salient team identity (like ours), our design does not exclude these potential interaction effects. We discuss these and other aspects related to differences across treatments further in Section 1.4.

1.3 Results

1.3.1 Team performance

We employ two outcome variables to measure team performance. First, to capture effects on the extensive margin, we consider whether a team manages to complete the task within the given time limit of 60 minutes. Second, we consider variation on the intensive margin by studying teams' finishing times, i.e., the time needed to complete the task.[17] Our main analyses focuses on capturing the effects of introducing the three distinct components of a tournament, *Identity*, *Rank*, and *Prize*. That is, we focus on comparing each "subsequent" treatment group to the "prior" one. To do so, we code a dummy variable for each component based on whether this component existed in the treatment the observation stems from. For example, in treatment *T2 (Identity, Rank)*, the dummy "*Identity*" and "*Rank*" are equal to 1, whereas the dummy *Prize* is equal to 0. This coding allows us to cleanly identify the effect of introducing the respective component (as compared

[17] Table 1.A.1 shows summary statistics of the probability of completion, finishing time, number of hints and the probability of purchasing a voucher by treatment.

Table 1.2: Team performance (completion and finishing time)

	Completed within 60 minutes				Finishing time			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
+ Identity	-0.086	-0.099	-0.048	-0.045	1.377	1.910	1.668	1.590
(making identity salient)	(0.052)	(0.066)	(0.060)	(0.056)	(0.870)	(0.970)	(1.180)	(1.117)
	[0.198]	[0.241]	[0.434]	[0.447]	[0.219]	[0.145]	[0.206]	[0.218]
+ Rank	0.105	0.093	0.081	0.079	-2.788*	-2.583*	-2.575**	-2.515**
(adding a ranking)	(0.046)	(0.049)	(0.048)	(0.045)	(0.856)	(0.801)	(0.851)	(0.836)
	[0.126]	[0.182]	[0.230]	[0.188]	[0.055]	[0.051]	[0.034]	[0.034]
+ Prize	0.091**	0.092**	0.079**	0.084**	-2.214**	-2.391**	-2.200**	-2.330*
(adding a prize)	(0.031)	(0.028)	(0.030)	(0.026)	(1.047)	(1.224)	(1.275)	(1.319)
	[0.047]	[0.032]	[0.033]	[0.020]	[0.042]	[0.033]	[0.040]	[0.064]
Mean in Control	0.527	0.527	0.527	0.527	56.470	56.470	56.470	56.470
Observations	373	373	373	373	373	373	373	373
Team Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Staff FE	No	No	Yes	Yes	No	No	Yes	Yes
Weekday 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)). The main explanatory variables are indicators whether the observation stems from a treatment that included the component(s) *Identity*, *Rank*, or *Prize*. All columns include room fixed effects. Each column indicates whether team controls (group size, share of males, experience, median age, language, private), staff, and weekday fixed effects are included. Standard errors in parentheses are clustered at the week level. p-values from score bootstrapping following Kline and Santos (2012) are listed in square brackets, with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

to the "prior" condition) on our outcome measures.[18] The results are shown in Table 1.2, and all specifications include room fixed effects to take into account the differing levels of difficulty that each room bears. We cluster standard errors at the weekly level (the level of treatment assignment), and, because of the relatively low number of clusters, we provide p-values from score bootstrapping following Kline and Santos (2012).

Columns (1) through (4) of Table 1.2 provide results from a series of Probit regressions, in which we estimate the marginal effects of each component on the probability of successfully completing the task. We control for team characteristics starting in Column (2), and add fixed effects for the ETR staff member on duty from Column (3). Column (4) shows our preferred specification, which also includes a fixed effect for the day of the week. Columns (5) through (8) repeat the same step-wise inclusion of controls and

[18] In Appendix Section 1.A.2, we provide results from additional analyses in which we use treatment dummies instead. These are in line with the results presented in the main text.

fixed effects, but instead use the time a team needs to complete the task as the dependent variable in a series of Tobit regressions (with 60 minutes as the upper limit).

The top row shows the results from making the identity salient. Counter our expectations, teams in treatments encompassing the component *Identity* are not more likely to complete the task in 60 minutes, nor do they finish earlier than in *Control*. The coefficients are statistically insignificant and, if anything, teams in *Identity* were less successful than teams in *Control*. Finally, the effect sizes of *Identity* are of relatively small magnitudes as compared to the effects of the other components when controlling for weekday fixed effects (Columns (4) and (8)). We conclude with Result 1:

Result 1 *Salient identity alone does not improve team performance.*

Adding a ranking (on top of making participants choose a team name) tends to make teams more likely to complete the task within 60 minutes (see Columns (1) through (4)) but the results are statistically insignificant due to the relatively large standard errors. However, adding a ranking significantly improves teams' finishing times by about 2.5 minutes (see Columns (5) through (8)). Hence, image concerns mainly enhance performance at the intensive margin (in line with the idea that mostly top performing teams are affected). We summarize these findings in Result 2:

Result 2 *Adding a weekly competition for social image improves team performance along the intensive, but not significantly so along the extensive margin.*

Adding a *Prize* to the weekly competition results in statistically significant performance improvements (see bottom row of Table 1.2). Teams are approximately 8 percentage points more likely to successfully complete the task within the time frame, and require 2.3 minutes less for completion. We conclude with Result 3:

Result 3 *Adding a prize to the weekly competition improves team performance along the extensive and intensive margins.*

As has become clear, we have found that tournaments can effectively improve team performance in non-routine tasks. Overall, the tournament with a prize increases the completion rate by more than 20 percent (almost 12 percentage points) and reduces finishing times by more than 3 minutes (remaining times are almost doubled, see also Table

1.A.2). Additional robustness tests for our main results can be found in the Appendix. Section 1.A.2 provides analyses based on treatment dummies instead of a component-based approach, with similar results. In Section 1.A.3, we conduct a randomization inference exercise confirming our findings.

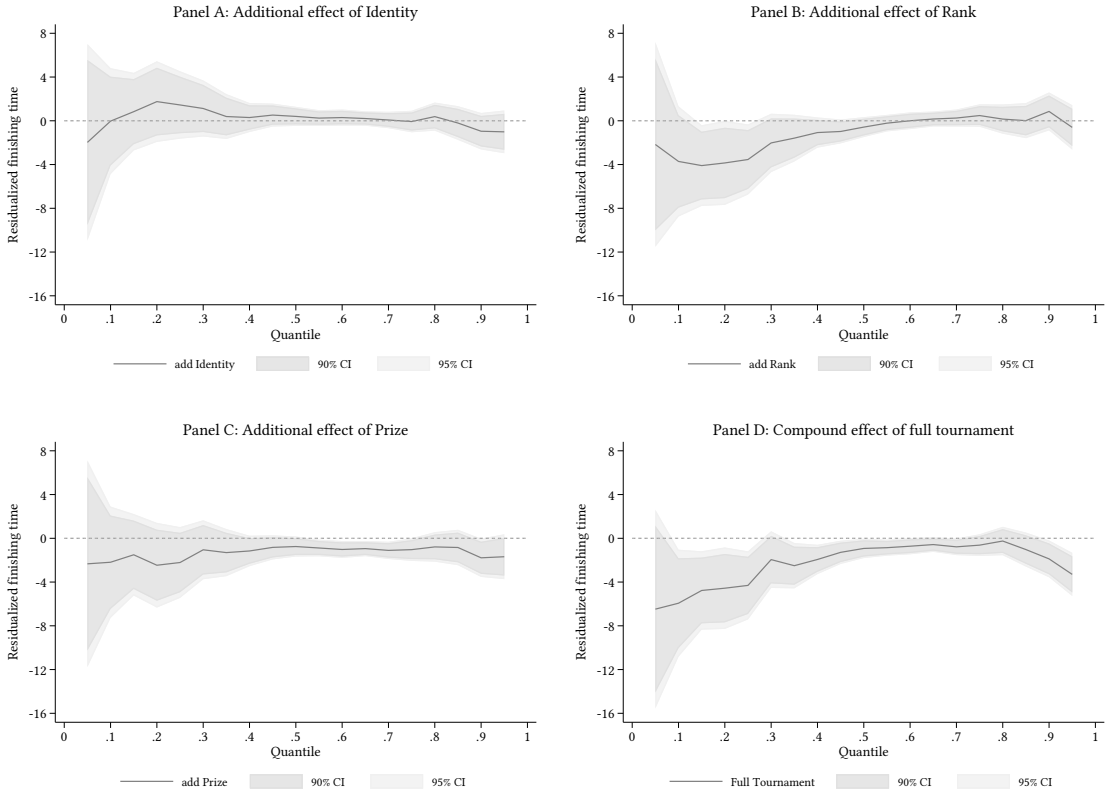
1.3.2 Team characteristics and the efficacy of tournaments

Competition for ranks and prizes may affect teams differently, due to their composition and potential for performance. To investigate such heterogeneity, we begin by illustrating in more detail how ranks and prizes influence teams across the entire performance spectrum using quantile regressions on residualized finishing times. We predict finishing times and residuals for all teams using the same fully specified Tobit regression as in Table 1.2, Column (8), including team controls, room, staff and weekday fixed effects.[19]

Panel A of Figure 1.1 shows that asking teams to discuss and choose a team name jointly before working on the task does not affect performance along the whole performance distribution (confirming Result 1). Panel B of Figure 1.1 shows that adding a weekly competition with a public ranking to a setting in which teams jointly deliberate on team names reduced the finishing times of the top performers, i.e., the lowest quantiles. This extra effect of rank incentives on the residualized finishing times declines along the performance distribution and becomes indistinguishable from zero around the 30% quantile. Panel C compares the residualized finishing times of teams being ranked and additionally eligible for a monetary prize with those of teams that are ranked but not eligible for a prize. Three interesting findings arise. First, adding a prize seems to further improve the finishing times of top performers substantially, but the effect lacks statistical significance due to the large confidence bands. Second, the positive impact of monetary prizes over rankings becomes significant around the 50% quantile and turns insignificant beyond the 75% quantile. Third, even though not always statistically significant, the estimated effects of adding a prize are all of similar magnitudes across the quantiles, suggesting a positive effect on the entire performance distribution. Panel D shows a comparison of residualized finishing times between *Control* and *T3 (Identity, Rank, Prize)*, and thus the compound effect of implementing the full tournament including the ranking with team

[19]The results in Table 1.2 did not show any performance improvement of *Identity* over *Control*. To increase statistical power, we therefore use observations from both *Identity* and *Control* for predicting finishing times. Using GLM (instead of Tobit) yields similar results (see Figure 1.A.3).

Figure 1.1: Quantile regressions on residualized finishing times



Notes: The figure shows quantile regressions on residualized finishing times. Panel A shows the additional effect of salient team identity. Panel B shows the additional effect of a public ranking. Panel C shows the additional effect of a monetary prize. And Panel D shows the overall effect of a tournament with a monetary prize (compares *T3 (Identity, Rank, Prize)* to *Control*). The line at zero marks residualized finishing times in the comparison group. Negative (positive) values indicate reductions (increases) in residualized finishing times due to adding component *Identity* (Panel A), *Rank* (Panel B), and *Prize* (Panel C), or due to adding all tournament features simultaneously (Panel D).

names and the monetary prize. This tournament improves performance along a large part of the distribution, so that teams facing salient team identity, image, and instrumental concerns perform better than similarly composed teams under the *Control* condition. In settings where top performance is particularly important, such as in many innovation contexts, public rankings, therefore, seem to be highly effective, whereas monetary prizes may additionally stimulate performance also below the very top.

In additional exploratory analyses, we also study possible heterogeneity in the observed treatment effects. To do so, we conducted additional regression analyses including interaction terms between each treatment component (i.e., the dummy variables *Identity*, *Rank*, and *Prize*) and observable team characteristics, presented in Appendix Sec-

tion 1.A.4. We do not find strong heterogeneity in the efficacy of our treatments, but suggestive evidence that rankings are particularly effective when teams are mostly composed of men (in line with the previous literature on competition and gender in routine tasks (Niederle and Vesterlund, 2011; Schram et al., 2019)).

1.3.3 Willingness to explore original solutions and potential crowding out

Prior research has suggested that incentives and competition may be ineffective (or even counterproductive) when production involves non-routine tasks that require thinking out of the box. Incentives may lead to focusing (Duncker, 1945), and thereby reduce thinking out of the box, and, in complex tasks, incentives may systematically discourage the exploration of new and original approaches (e.g. Amabile, 1996; Azoulay et al., 2011; Ederer and Manso, 2013; McCullers, 1978; McGraw, 1978). Furthermore, we incentivized performance in terms of teams' finishing times and not according to the willingness to explore original solutions on their own. Teams may thus substitute speed for such exploration, particularly when they face difficult problems (for an excellent discussion and evidence from the laboratory see also Laske and Schroeder, 2016).

Our setting offers the possibility of testing for such potential discouragement or substitution, as teams had the opportunity to seek external help using up to five hints, which did not negatively affect their rank in the tournament. We focus on the number of hints taken as well as on the timing of the hints. In general, the number of hints and finishing times are positively correlated (Spearman's $\rho = 0.5191$, $p < 0.001$), as worse teams are on average more likely to seek help by taking hints. However, the number of hints requested does not differ significantly across treatments (Kruskal–Wallis test, $p = 0.6041$). If anything, it appears as if teams exposed to component *Prize* take on average slightly fewer hints (average number of hints taken in *Control*: 3.39, *T1 (Identity)*: 3.29, *T2 (Identity, Rank)*: 3.25, *T3 (Identity, Rank, Prize)*: 3.18). Additional analyses confirm that despite the positive effect on performance, the addition of none of the components (*Identity*, *Rank*, or *Prize*) significantly increases the number of hints taken, nor their timing (see Appendix Section 1.A.5). These results indicate that work environments in innovation contexts sharing the features of our team task are unlikely susceptible to a reduction in teams' inclination to explore own approaches due to tournaments.

Offering extrinsic incentives could also crowd out intrinsic motivation (e.g. Deci et al., 1999; Eckartz et al., 2012; Gerhart and Fang, 2015; Hennessey and Amabile, 2010) to perform the task at all. The challenging nature of non-routine analytical tasks renders them particularly exciting for intrinsically motivated workers (for a discussion see also Autor and Handel, 2013; Delfgaauw and Dur, 2010; Friebel and Giannetti, 2009), as in these settings workers can make new discoveries and experience progress jointly. Our setting provides us with teams that are highly motivated to perform the task (teams are even willing to pay for facing the challenge) and thus a unique opportunity to test whether image and instrumental concerns innate to tournaments affect the intrinsic motivation to perform a similar task again. To evaluate whether the addition of any component indeed reduced a team's intrinsic motivation, we focus on a revealed preference measure. After completion of the task, all teams were offered the opportunity to buy a voucher at a reduced price allowing them to perform a new but comparable task again (at any branch of *ExitTheRoom*).

In contrast to the idea that tournaments may reduce a team's intrinsic motivation to work on a similar task again, we find small, positive, but statistically insignificant effects (see Appendix Table 1.A.8). As such, our findings speak against a substantial crowding out of intrinsic motivation for future participation and underline the positive roles of image and instrumental concerns innate to tournaments.

1.4 Discussion

Our experimental treatments step-wise introduced three components innate to tournaments: salient team identity, image concerns, and instrumental concerns. However, we could not independently vary each of the three components as our collaboration partner considered some treatments resulting from a full factorial design incongruous (see also Section 1.2.2). Being constrained to the three implemented treatment conditions comes with the caveat that we cannot explicitly study potential interactions between the three different components (e.g., we cannot directly measure potential identity-enhancing effects of the introduction of a public ranking) and requires a more detailed discussion of which other potentially relevant changes each treatment variation may bring about.

As compared to the *Control* condition, *T1 (Identity)* ensures salient team identity, but the limited treatment effect may have eventually resulted from team identity being also

salient in *Control* (as about 80% of teams were composed of friends). This aspect leaves important room for future studies on the role of identity for team performance in non-routine tasks but renders potential additional identity enhancing effects of team competitions (by adding a ranking and a prize) in our setting less likely. Further, it is plausible that the introduction of public rankings may not only result in image concerns but also render time to completion a more relevant performance outcome. Similarly, adding a prize to the competition for ranks may not only introduce instrumental concerns but additionally render the role of finishing times salient, and such shifts in focus may improve team performance independently of image and instrumental concerns. Importantly, Englmaier et al. (2018, p.22) show that a focus on finishing times alone does not improve performance in escape challenges in the exact same setting, such that the observed performance improvements due to the introduction of the competition for ranks very likely result from additional image concerns, rather than from an interaction with salient team identity. Finally, introducing a prize may not only result in instrumental concerns but also alter image concerns. Teams may perceive the value of appearing first in the public ranking differently, because monetary prizes may either crowd out parts of the image motivation or increase the image value of being ranked high. As we observe that the introduction of ranks particularly boosts performance of teams at the top of the performance distribution while introducing a prize leads to improvements along the whole performance spectrum, we consider it less likely that the addition of a monetary reward substantially altered image concerns which then caused the observed performance improvements.

1.5 Conclusion

Tournaments are an important and often-used mechanism to foster innovation (Lindegaard, 2010; Terwiesch and Ulrich, 2009; Terwiesch and Xu, 2008; Scotchmer, 2004). They not only involve instrumental incentives but also include important behavioral aspects that can foster team performance in non-routine tasks. Our study exploited the unique opportunity to exogenously vary features innate to typical tournament incentives (salient team identity, team rankings, and prizes) treating a large number of teams performing a non-routine analytical task in a natural field experiment. We found that fully-fledged tournament incentives, in which teams compete for a monetary prize awarded to the best

performing team listed in a public ranking of team names, substantially improved team performance. Public rankings of team names alone improved performances of teams expected to be at the top of the performance distribution but did not affect teams at the bottom. Lastly, rendering team identity salient by having teams jointly deliberate on their team name (see Ai et al., 2022) was not enough to improve performance on its own.

Complementing this novel field-experimental evidence on the effects of tournaments for team performance in non-routine tasks, we further showed that performance improvements due to tournaments did not result in a reduction of teams' willingness to explore solutions on their own. Further, we found no indications of a reduction of teams' intrinsic motivation to perform similar tasks again in the future due to tournament incentives. As we elicited a revealed preference measure of a team's willingness to work on a similar task before the team receives actual feedback on its relative performance, this finding suggests that potentially negative effects of rank or tournament incentives observed in routine tasks (see e.g. Barankay, 2012; Ashraf et al., 2014; Ashraf, 2019; Blader et al., 2020) likely result from actual, discouraging performance feedback for underperforming teams rather than from the anticipation of such feedback or competition per se. Avoiding such feedback, we thus found robust evidence for the important roles of image and instrumental concerns in the efficacy of tournaments in non-routine analytical team tasks.

Overall, our results make an important contribution to the literature on teamwork in non-routine analytical tasks with a clearly specified goal and deadline. We confirm and extend findings from laboratory experiments on closed-form creativity (Charness et al., 2014; Charness and Grieco, 2019) and show that tournaments can substantially improve performance in a novel and challenging field setting. Thereby, we provide basis for important future field work. One fruitful avenue for such research lies in studying whether image and instrumental concerns lead to adjustments in team organization. For example, Englmaier et al. (2018) find suggestive evidence that bonus incentives can alter team organization in the same setting and are accompanied by an increased demand for leadership. Following these results, it will be interesting to investigate whether tournament incentives and leadership are substitutes or complements. Further, it will be interesting to investigate the role competitions with and without prizes in field settings with open-form tasks.

Chapter 2

The Value of Leadership: Evidence From a Large Scale Field Experiment [1]

Abstract

Companies increasingly make use of team-based organizational structures. To foster performance in these settings, scholars and practitioners alike have emphasized the potential of leadership. However, the causal impact of leadership in agile and cross-functional teams is difficult to identify since leadership is often determined endogenously. In a large-scale natural field experiment (1,273 participants, 281 teams), we randomly encourage teams to select a leader before performing a complex non-routine, analytical task. This encouragement substantially increases the fraction of teams completing the task and makes successful teams faster. Choosing a leader also improves team organization, without affecting the originality of solutions.

[1] This chapter is based on joint work with Florian Englmaier (LMU Munich), Stefan Grimm (LMU Munich), David Schindler (Tilburg University) and Simeon Schudy (LMU Munich).

2.1 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 often face complex tasks that require team members to exert cognitive effort, stay motivated, and work in a coordinated manner. Thus, teams may benefit not only 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., 2022), 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 the presence of leadership is often determined endogenously. Consequently, causal estimates of the efficacy of endogenous leadership are largely missing.[2] This study exploits a unique opportunity to uncover the causal effects of the presence of endogenously chosen leaders for team performance in a non-routine team task. To exogenously vary the presence of 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 real-life escape challenges. This setting encompasses important elements encountered in many other non-routine, analytical, and interactive team tasks and is nowadays also used to recruit high-skilled workers as well as to assess and improve individuals' teamwork ability and leadership skills.[3] Escape challenges provide a unique environment to study the value of leadership in non-routine

[2] See, for example, the meta-analysis on shared leadership by Nicolaides et al. (2014, p. 936), which relies on correlational evidence.

[3] 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).

tasks. First, teams must 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 determination of a leader. Third, teams encounter problems that are novel and challenging for them but kept identical across teams and are 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 challenge provider allows us to randomly assign experimental treatment conditions to many teams that are unaware they are 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 challenge provider ExitTheRoom (ETR), 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 emphasizes the positive role of leadership before teamwork starts but does not enforce the choice of a leader. This simple variation allows us to identify the value of leadership encouragement 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 they 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 the wedge between the 60 minutes time available for solving the task and a team's actual finishing time (i.e., a team's 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 yield similarly positive effects on team performance and 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 work together on sub-tasks. Hence, leadership seems to increase decentralized information acquisition and problem solving. As leadership changes team organization and results in performance increases, it likely improves coordination among team members. This latter interpretation is also reflected in teams' perceptions of coordination, which we were allowed to elicit as part of a short customer survey after the escape challenge.

In addition, our setting allows us to consider potential impacts of *Leadership* on the willingness to explore original solutions. During the escape challenge, teams can 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' propensity to provide original solutions (see also Englmaier et al., 2018), we find that *Leadership* does not affect teams' willingness to explore original 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 the first field evidence on the causal effect of leadership encouragement in teams that may endogenously choose a leader when performing a non-routine task. In contrast to important field work that has studied the causal effects of exogenously assigning a leader (Boudreau et al., 2021) or different leadership styles (Kvaløy et al., 2015; Meslec et al., 2020; Antonakis et al., 2022), we focus on the value of choosing a leader. That is, we do not compare the quality of leadership, the leadership style of bosses (Bertrand and Schoar, 2003; Lazear et al., 2015; Bandiera et al., 2020; Bennedsen et al., 2020; Hoffman and Tadelis, 2021), or how different management practices impact productivity (see, e.g., Bloom and Van Reenen, 2007; Bloom et al., 2013; Bruhn et al., 2018; Gosnell et al., 2020). Instead, we provide an estimate of the value of leadership itself.[4] Further, our study is unique in focusing on the value of leadership for teamwork in a non-routine analytical task rather than on individual performance in routine tasks (Kvaløy et al., 2015; Antonakis et al., 2022; Meslec et al., 2020). Additionally, and related to a large body of laboratory experimental evidence on the positive effects of leadership on coordination (e.g., Weber

[4]For an interesting theoretical argument on the relative value of leadership as compared to flat hierarchies in teams see also Dessein (2007).

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 show that leadership can also alter team organization and improve (perceived) coordination among team members 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).[5] Other work in this domain focuses on the role of monetary incentives for idea creation and team performance (see, e.g., Gibbs et al., 2017; Englmaier et al., 2018, 2023a), and finds positive incentive effects. Most closely related to our setting, Englmaier et al. (2018) study the effect of offering a monetary bonus of 50 Euros for completing an escape challenge within 45 minutes instead of 60 and find that the bonus increases 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. We thus identify a substantial value of leadership 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 encouragement thus appears as a cost-effective tool to foster team performance. We find that leadership may help to efficiently delegate individual sub-tasks without hampering teams' ability to efficiently master the challenge they face. Hence, to foster joint production in agile and cross-functional teams, companies may substantially benefit from emphasizing the important role of an overall project leader before teamwork begins.

The rest of this paper is structured as follows. Section 2.2 describes our experimental design, measurements, and procedures in more detail. We provide results from the experiment in Section 2.3. Section 2.4 investigates potential mechanisms, and Section 2.5 concludes.

[5]These tasks include activities that involve cognitive rather than physical effort, are interpersonal, and involve forming and testing hypotheses. More broadly, they also include forms of creative production (see, e.g., Ramm et al., 2013; Bradler et al., 2019; Charness and Grieco, 2019; Gibbs et al., 2017; Laske and Schroeder, 2016).

2.2 Experimental design

2.2.1 The field setting

We collaborate with ETR, a provider of real-life escape challenges.[6] In escape challenges, teams of customers are confronted with a cognitively demanding team challenge that is non-routine and interactive. The goal is to complete the challenge within a limited amount of time (60 minutes). The challenge is composed of a series of quests that ultimately yield a final code to solve the task and succeed. To complete the task, teams must search for clues, combine the collected information, and think outside the box. They also often need to make unusual use of objects and develop and exchange innovative ideas to arrive at the solution. If the team manages to succeed before the 60 minutes expire, they win, and if time runs out before the team solves all quests, they lose. To maximize the chances of completing the task, teams have a strong incentive to succeed as fast as possible, as they do not know how many quests have to be solved in total.[7]

Escape challenges 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. We conducted our experiments at ETR's facilities in Munich, Germany. The location offers three challenges with different themes and background stories.[8] Teams have a time limit of 60 minutes, and the remaining time is displayed at all times in the rooms. If they get stuck, they can request up to five hints (they must state explicitly that

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

[7] More generally, making progress together by solving complex problems is at the core of these team challenges and teams' strong motivation to finish quickly is also reflected by teams proudly writing their finishing times on the walls of the entrance area of ETR. Englmaier et al. (2018, pp. 6-7) provide an example of a typical quest in a real-life escape challenge to illustrate the nature of the task in more detail. Escape challenges are usually 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. We present this example here as well since our partner asked us not to reveal actual content. In the (fictitious) example, a team has found several objects in a room, among them an unlocked box that contains a megaphone, which can be used as a speaker and can also play three distinct types of alarm sounds. There is also a volume unit (VU) meter in one corner of the room. To open a padlock on a box containing additional information, the team needs a three-digit code. They obtain this code by playing the three types of alarms on the megaphone and writing down the corresponding readings from the VU meter. The teams at ETR solve quests similar to this fictitious example. These quests may further include finding hidden information in pictures, forming and testing hypotheses by combining different pieces of information, constructing objects (e.g. a flashlight) out of several parts, or identifying and solving rebus (word picture) puzzles (see also Erat and Gneezy, 2016; Kachelmaier et al., 2008).

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

they need help) via a walkie-talkie from ETR staff. These hints never include the direct solution but only provide vague clues regarding the next required step.

2.2.2 Experimental treatments and procedures

We pre-registered the experimental design with the AEA registry (AEARCTR-0002570) and conducted our experiment at ETR between January and March 2018 during their regular opening hours from Monday to Thursday. The 1,273 participants in 281 teams were all regular ETR customers. Teams booked specific time slots through ETR's website, usually several days in advance. Upon arrival, staff welcomed the teams and asked them to sign ETR's terms and conditions, including its data privacy policy. The staff then delivered a standardized introduction including the narrative of the booked event and the general rules at ETR, and they guided the 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 to avoid treatment spillovers between different teams on-site (due to the chance of meeting other participants when arriving early for a subsequent slot).[9] In the *Control* treatment (95 teams), staff welcomed teams without further intervention. In the *Leadership* treatment, staff welcomed the teams, highlighted the importance of leadership to succeed in the task, and encouraged them to select a leader according to a short standardized script (see below). To more closely investigate the effects of different types of leadership (see, e.g., Bass, 1999), *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 stressed the leader's role differently, 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."[10]

[9] 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). Table 2.A.1 excludes these cases and shows that our main conclusions do not hinge on the inclusion of these observations.

[10] 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 the *Motivation* treatment,

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 *Control* condition, we pool the data for the main analyses and use both sub-treatments when discussing mechanisms.

2.2.3 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, the team's gender and age composition, the language the team spoke (German or English), experience with escape challenges, and whether the customers came as a private group or were part of a company team-building event.[11] Additionally, as a proxy for teams' propensity to have someone take the lead, we collected information about whether one team member took 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 noted whether team members searched for information individually (as opposed to jointly) and whether teams were spending much time working together (versus spread out across the room) on a five-point Likert scale (from 1 = "not at all" to 5 = "a lot").[12]

Table 2.1 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

ETR staff mentions the word "motivates," while in the *Coordination* treatment they mention "coordinates." Naturally, the statement also relates to performance and mentions the term "team" explicitly. These terms may render team identity more salient and shift focus on performance. However, neither making team identity more salient (Englmaier et al., 2023a) nor a focus shift on performance (Englmaier et al., 2018) improves performance in the same team challenge.

[11] 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. 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 person's 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.

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

Table 2.1: 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 following 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 are in parentheses; (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 *** = $p < 0.01$.

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 experiencing the escape challenge. This survey included questions on overall satisfaction with the team challenge, the value for money, exerted effort level, and perceived coordination and motivation in the team. All questions were answered on an eight-point Likert scale.

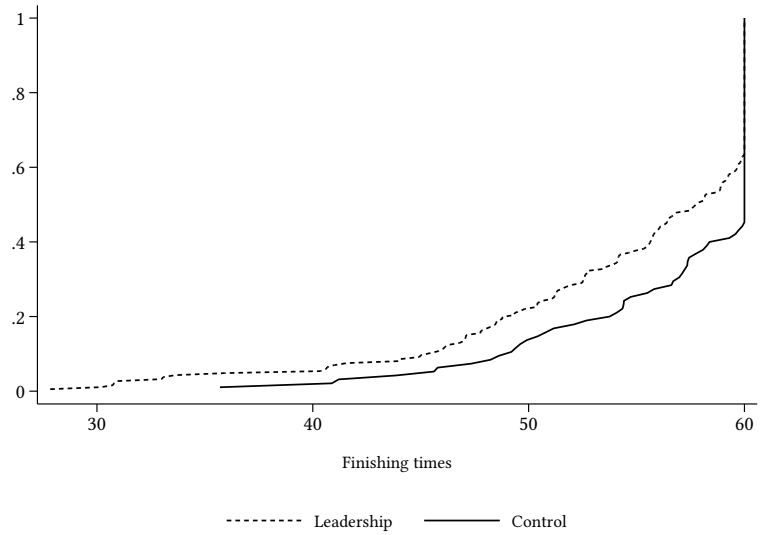
2.3 Results

2.3.1 Team performance

Figure 2.1 shows the cumulative distribution functions (CDFs) of finishing times across conditions. Teams in the *Leadership* treatment condition clearly perform better than those in the *Control* condition. Specifically, 63% of teams finish the task within the time limit of 60 minutes in *Leadership*, whereas only around 44% do so in *Control* (Pearson χ^2 test: $p < 0.01$). In addition to being more likely to complete the task, teams that were encouraged to choose a leader also solve the task faster (average remaining times: 3m10s in *Control*, 5m29s in *Leadership*, Mann-Whitney test: $p < 0.01$).

These non-parametric results are confirmed by a series of Probit regressions, in which we step-wise introduce additional control variables. To account for differences in the task teams face, all specifications include room fixed effects. In Column (1) of Table 2.2, we

Figure 2.1: CDFs of finishing time



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

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 2.1). To account for potentially idiosyncratic behavior by ETR staff who delivered the general instructions and leadership encouragement, we employ staff member (including our own research 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 CDFs of finishing times in *Leadership* and *Control* (see Figure 2.1) indicate 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* first-order stochastically dominates the CDF of *Leadership*, and the data skew toward the end and are very flat in the left tail. Further, finishing times are censored at 60 minutes.

Table 2.2: Team performance (completion and finishing time)

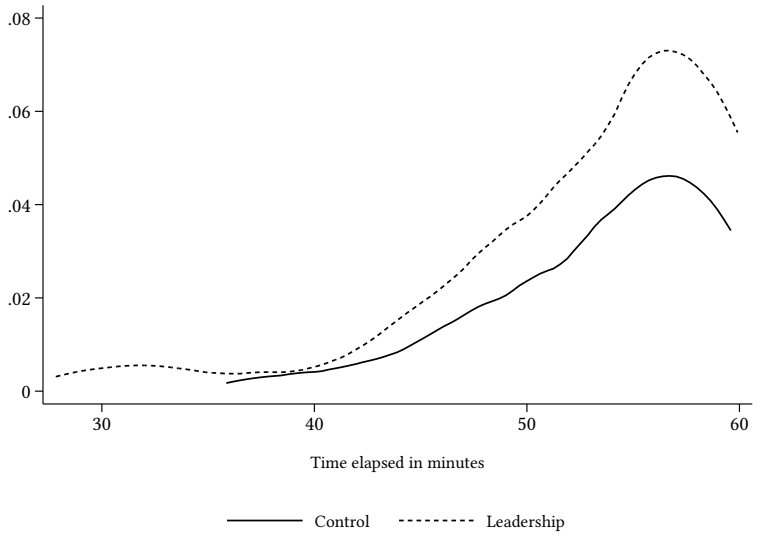
	Completed within 60 Minutes				Finishing Time			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Leadership	0.137***	0.137***	0.125**	0.108**	-3.175***	-3.037***	-2.773**	-2.551**
	(0.045)	(0.047)	(0.058)	(0.043)	(0.912)	(0.873)	(1.137)	(1.253)
Mean in Control	0.442	0.442	0.442	0.442	56.814	56.814	56.814	56.814
Observations	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)–(4)) and Tobit regressions of finishing time (Columns (5)–(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 *** = $p < 0.01$.

To avoid underestimating the treatment effect and to 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). Columns (5) through (8) of Table 2.2 reveal a statistically significant and sizable reduction of finishing times in *Leadership* in all four specifications. Teams are, on average, two-and-a-half minutes faster, which is equivalent to an increase of about 75% of teams' remaining times.

Finally, Figure 2.2 provides the results from a hazard model (survival analyses) in which finishing the task is considered the "hazard." The figure illustrates hazard rates of completing the task, conditional on not yet having it completed, separately for both conditions. It 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. We do not find that leadership substantially improved team performance at the lower end of the performance distribution.

Figure 2.2: Hazard rates of finishing the task



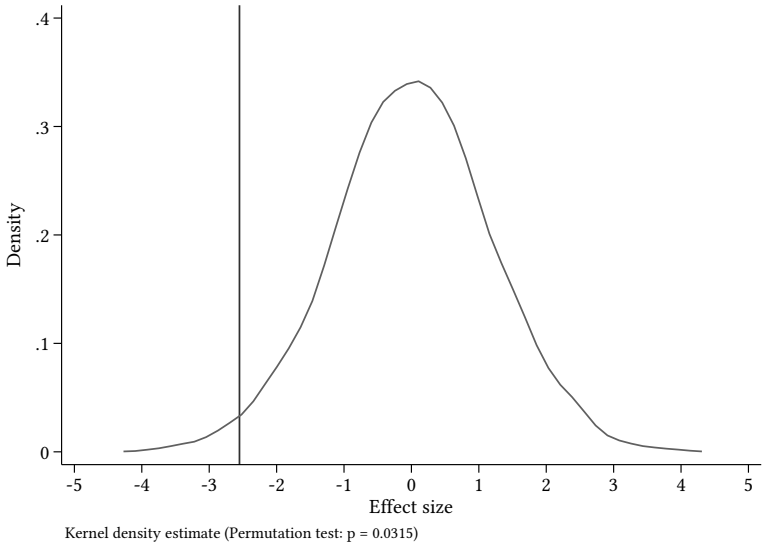
Notes: The figure shows the hazard rates of finishing the task (conditional on not having finished yet) separately for teams we randomly encouraged to select a leader (*Leadership*) and teams in the *Control* condition.

2.3.2 Robustness

To explore the robustness of our estimates, 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 counterfactual. We repeat this procedure 10,000 times, generating a distribution of counterfactual estimates we can compare to our "true" estimate. Figure 2.3 plots the distributions for teams' finishing times. The kernel density estimate is centered at zero 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 are "extreme" such that we can confidentially reject the null hypothesis of no effect of our actual treatment (p-value $= 0.0315$).

Further robustness analyses are relegated to the Appendix. Appendix Table 2.A.1 repeats the specifications from Table 2.2 but excludes 12 teams, for which ETR staff did not implement the randomly assigned treatment correctly. Our conclusions remain unaffected. Appendix Table 2.A.2 shows the 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

Figure 2.3: 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.

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 2.2. Further, we study heterogeneity in reactions to the treatment. Figure 2.A.1 sheds light on whether teams in corporate bookings react differently to the treatment than teams in private bookings. Both, private and corporate teams, and thus also teams that frequently work together in a business context, benefit similarly from *Leadership* (see Appendix Tables 2.A.3 and 2.A.4). Tables 2.A.3 and 2.A.4 also show that there are no strong differences in the efficacy of *Leadership* based on other underlying team characteristics (e.g., teams with or without prior experience).[13]

[13]Only 1 out of the 14 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 5% level. The result should, however, be taken with a grain of salt, as only a small minority of teams does not speak German.

Table 2.3: 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 *** = $p < 0.01$.

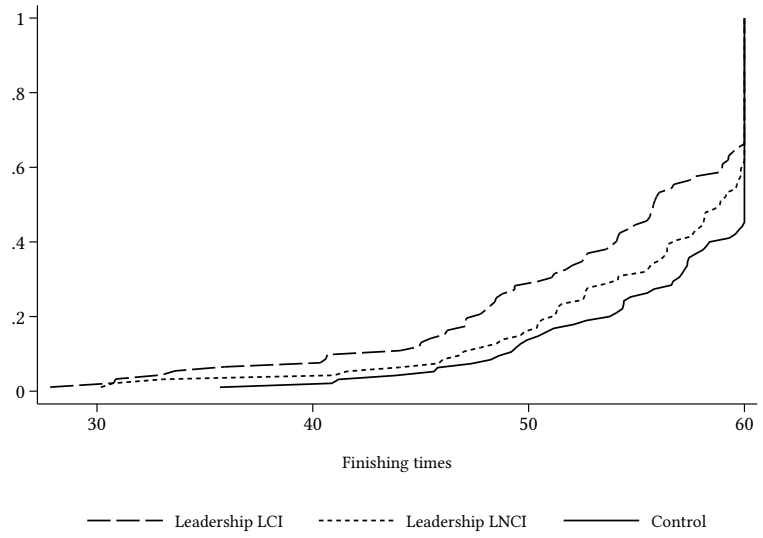
2.4 Mechanisms

2.4.1 The framing of leadership functions

As described in Section 2.2.2, we framed the role of leaders differently in the two sub-treatments, *Motivation* and *Coordination*. In *Motivation*, we suggested that the group may want to choose a leader "...who takes over the leading role and consistently motivates the team", while in sub-treatment *Coordination*, the leader was supposed to "...consistently coordinate the team." In Table 2.3, 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.[14]

[14] 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 can expand on the overall effect of leadership we detected.

Figure 2.4: CDFs of finishing times



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

2.4.2 Choosing a leader

Next, we investigate whether those teams that actually chose a leader also perform better. Around 50% of the teams encouraged to choose a leader do so before working on the task, whereas we did not observe a single team explicitly choosing a leader in *Control* before teamwork began. Regression analyses in Appendix Table 2.A.5 further indicate that the immediate choice of a leader does not relate systematically to observable team characteristics.[15]

As choosing a leader is equally likely in both sub-treatments (see Appendix Table 2.A.5, Column (2)), we again focus on our main treatment condition *Leadership*. Figure 2.4 shows the CDFs of finishing times in *Leadership* depending on whether a leader was chosen immediately (LCI) or not chosen immediately (LNCI) as well as finishing times

[15]Similarly, as shown in Appendix Table 2.A.5, Column (3), observable team characteristics have limited predictive power for the chosen leader's gender (fewer 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 leader's 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.

Table 2.4: 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): we first 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. We then 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 *** = $p < 0.01$.

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 versus *Control*, $p < 0.01$; LNCI versus *Control*, $p < 0.10$) and the extensive margin (Pearson χ^2: LCI versus *Control*, $p < 0.01$; LNCI versus *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 versus LNCI, $p = 0.09$) but less so at the extensive margin (Pearson χ^2: LCI versus LNCI, $p = 0.51$).[16]

To analyze whether teams that immediately chose a leader were more successful, we follow the procedure recommended in Angrist and Pischke (2008, p. 142) and employ a two-stage approach. In the 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 the second step, we use these non-linear fitted values as instruments and estimate their impact on team performance. Table 2.4 presents the

[16]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 that never chose a leader and teams that chose a leader at a later stage while performing the task.

results from OLS and 2SLS regressions for comparison. Panel A reports the intention-to-treat (ITT) 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 2SLS results of the second stage. Further, 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 ITT 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, the results in Column (2) indicate that the coefficient of immediately choosing a leader in Panel B is larger than the ITT estimate in Panel A, indicating that teams choosing a leader immediately are indeed more successful and solve the task substantially faster.

2.4.3 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 team members' tendency to meet or work in close proximity (standing together) and their tendency to search individually for new information (individual search). Acquiring information individually may be beneficial if the team is well organized and exchanges the collected information, while working together may indicate joint acquisition or reflection on ideas, which may be less relevant when teams are well organized. Table 2.5 shows estimates from OLS and 2SLS regressions (using to the same approach as in Table 2.4) for the relationship between *Leadership* or choosing a leader and a teams' (standardized) tendency to stand together and search individually for information. The ITT estimate in Column (1), Panel A shows that being assigned to the *Leadership* condition reduces team members' tendency to meet or work in close proximity (standing 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 increases teams' propensity to search individually; and even more so for teams that chose a leader (Column (2), Panel B). This suggests that our leadership encouragement is effective because it increases teams' tendency to choose a leader and thereby

Table 2.5: 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 work together and search individually on our treatment indicator (with *Control* as base category). All variables are standardized with mean zero and a standard deviation of one. For 2SLS, we follow the procedure outlined by Angrist and Pischke (2008): we first 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. We then 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 *** = $p < 0.01$.

changes teams' strategies on how to acquire and process information. As, overall, leadership results in a substantial performance increase, teams that 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 standard deviations better than teams in *Control* (see Appendix Table 2.A.6, Column (5), in which we use teams' responses to the short customer survey).

Finally, our setting also allows us to study whether leaders affect how much teams explore original solutions. Recall that in the task all teams can request up to five hints by contacting ETR staff using a walkie-talkie if they get stuck. In Table 2.6, 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 very minor influence on the timing of hints. We thus conclude that *Leadership* improves team performance without affecting the willingness to explore original solutions.

Table 2.6: 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	281	281	281	281	281
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 the number (1) and timing of hints requested (2) – (6) on our treatment indicator (with *Control* as base category). For 2SLS, we follow the procedure described by Angrist and Pischke (2008): we first 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. We then use these nonlinear fitted values as instruments. All columns include room fixed effects, team controls, staff, weekday, and week fixed effects. Standard errors in parentheses are clustered at the daily level, with significance levels * = $p < 0.10$, ** = $p < 0.05$, and *** = $p < 0.01$.

2.5 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 can be improved by a simple encouragement to choose a leader before teamwork begins. We conducted a large-scale natural field experiment (Harrison and List, 2004) with 281 teams performing an escape 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 both the extensive and intensive margin.

We find that 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 time remaining until the deadline is 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 as well as improved team organization, without affecting the willingness to explore original solutions.

Apart from immediate implications for cost-effective improvements of team performance through leadership encouragement in practice, these findings also highlight 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. Assessing this comparisons may become especially important as some companies may wish to have their leadership elected endogenously (as the German company Deutsche Telekom did in 2017).[17] 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.[18]

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., 2022), a fruitful avenue for future research lies in studying whether endogenous leadership in teams and team incentives are substitutes or complements. Finally, following theoretical arguments by Hermalin (1998) and Bolton et al. (2013), 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.

[17] `https://www.kom.de/medien/fuehrungskraefte-wahl-bei-der-telekom/`, in German.

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

Chapter 3

The (Mis)Perceived Determinants of Team Success in Non-Routine Analytical Team Tasks [1]

Abstract

Over the last decades, work tasks have become increasingly non-routine, complex, and analytical, leading to the widespread adoption of team-based organizational structures. To assemble productive teams and implement efficient governance structures, human resource (HR) experts need to form correct expectations about the most crucial determinants of team success. This study documents HR experts' perceptions (n=3,000) regarding the relative importance of various team composition dimensions and governance structures for performance in non-routine analytical tasks. Exploiting the unique opportunity to contrast expectations with actual performance data of 1,062 teams, we show that experts hold qualitatively accurate beliefs. However, they substantially underestimate the value of leadership. These patterns hold up in an additional general population sample (n=3,000). Furthermore, we document implicit biases against (particularly female) leadership, which partially depend on the respondent's own gender.

[1] This chapter is based on joint work with Florian Englmaier (LMU Munich), David Schindler (Tilburg University) and Simeon Schudy (LMU Munich).

3.1 Introduction

After rising importance over the last decades, teamwork has become ubiquitous to the modern economy (Driskell et al., 2018; Deming, 2017). For example, academic research across nearly all disciplines is conducted increasingly by teams (Wuchty et al., 2007), and also in many other domains, teamwork prevails: Teams perform tasks in health care (Hughes et al., 2016), aviation (Littlepage et al., 2016), sports (McEwan and Beauchamp, 2014), the military (Dalenberg et al., 2009), space travel (Salas et al., 2015), and (most importantly from an economics perspective) in many firms (Marks et al., 2001). As work tasks have become more non-routine, complex, and analytical (Autor et al., 2003; Autor and Price, 2013), teamwork is also viewed as the central building block to success in modern firms (Bandiera et al., 2013; Weidmann and Deming, 2021).[2] Consequently, businesses and scholars alike seek to form the right expectations about what renders team production successful.[3]

A major challenge for firms is to compose teams such that they can work productively and to implement governance structures that effectively foster team performance, rendering the role of human resources (HR) experts key. In many firms, these experts are responsible for selecting team members and establishing the rules and processes by which teams operate. In turn, their expectations regarding optimal team composition and governance structures can significantly impact team performance. However, HR experts face a difficult task when forming expectations about the relative importance of different team characteristics and team governance attributes. First, causal evidence on the role of team composition and optimal governance structures for non-routine tasks is still scarce and did so far not allow for systematic comparisons across task determinants.[4] Second,

[2] This is also reflected in employers' demand for new employees' ability to work in teams (see, e.g., the National Association of Colleges and Employers Survey NACE, 2022).

[3] Two recent examples from the National Aeronautics and Space Administration (NASA) and Google underline how much organizations care about forming the right expectations regarding the determinants of team success. NASA has started two projects focusing on optimal team composition for successful space missions (*NASA CREWS* and *TEAMSTaR*). These projects seek to help stakeholders forming the right expectation (by predicting how team composition will affect the team's social relationships and performance) and providing them with decision support systems. Research at Google (*Project Aristotle*) has helped the company to obtain informed expectations about determinants of team performance in the past and particularly emphasized the importance of optimal team governance structures.

[4] A few recent studies include work by Hoogendoorn et al. (2013), who provide causal estimates of the (positive) effects of gender diversity in teams of undergraduate students who start up a venture as part of their curriculum, and Englmaier et al. (2018, 2023a, 2021) who study the efficacy of bonus incentives, tournament incentives, and leadership in a team escape challenge.

even if HR experts could acquire relevant information on what successfully shapes team performance, they are not immune to bias (see, e.g., Kübler et al., 2018).[5] Experts' socioeconomic characteristics (such as their gender or age) as well as (mis)perceived social norms (i.e., experts' expectations about their peers judgments on related issues) may bias experts' expectations and eventually result in suboptimal decisions (see also Bursztyn et al., 2020). It is thus key to better understand how experts perceive the relative importance of different determinants of team success, whether they misperceive the empirical relevance of these determinants, and whether they are prone to additional (implicit) biases due to their own socioeconomic characteristics or (mis)perceived social norms.

This study provides a unique and comprehensive analysis regarding HR experts' perceptions of several key attributes relating to team composition and governance structures in non-routine analytical tasks and exploits the unique opportunity to contrast experts' perceptions with predictions based on actual performance data. To elicit experts' (n=3,000) perceptions, we use an incentive-compatible method by combining a discrete choice experiment with the incentivization method introduced by Bardsley (2000). In the experiment, experts face a series of discrete choices between two teams with varying attributes related to team composition and governance structures. Experts are incentivized to bet on the team they consider more successful, and know that they can receive a monetary reward of 100 € if their choice coincides with the team actually performing better (in one randomly selected actual team comparison). Experts are fully informed about the nature of the non-routine task teams perform and see a variety of combinations of key attributes related to team composition and governance structures across these comparisons.

There are many candidate attributes that may affect team performance. Relating to team composition, prior research has linked team size with team performance (Kozlowski and Bell, 2013; Stewart, 2006), highlighting that increasing team size may help teams (due to the availability of more resources) but also, that large teams may suffer from coordination failures and losses due to miscommunication. Further, there is an ongoing discussion on the value of diversity in teams. On the one hand, diversity promises large benefits for teamwork as more diverse teams may be more likely to come up with innovative ideas (Hoogendoorn et al., 2013). On the other hand, diversity may increase communication or

[5] Toma and Bell (2021) also provide evidence that choices are often inelastic to the provision of relevant information on impact.

coordination costs (Lyons, 2017).[6] Further, the literature suggests that expectations about the value of diversity may be prone to misperceptions. For example, Sarsons (2017), Isaksson (2018), Coffman et al. (2021), and Sarsons et al. (2021) show that women receive less credit for their joint work with men, and that the degree to which team members of different genders receive recognition for their work may depend on stereotypes. Finally, task-specific human capital of team members (e.g., whether they have already performed similar tasks) may matter for team performance (Bartel et al., 2014). Related to team governance structures, monetary incentives in the form of bonuses or tournaments as well as competitions for status are likely candidates that may affect performance: While incentives may discourage the exploration of new and original approaches in complex, non-routine tasks (e.g. Amabile, 1996; Azoulay et al., 2011; Ederer and Manso, 2013; McCullers, 1978; McGraw, 1978), they may also foster idea creation and team performance (Gibbs et al., 2017; Englmaier et al., 2018, 2023a).[7] Further, leadership has been attributed great importance in business, management, economics, and politics (Antonakis et al., 2022). Teams working on non-routine tasks may benefit from leaders who motivate (House, 1976; Bass, 1998, 1999; Howell and Avolio, 1993) or coordinate (Bass, 1990; House et al., 1999) their team members and thus affect team performance (Englmaier et al., 2021).

Our comprehensive approach elicits experts' perceptions concerning many of the above-mentioned attributes. With respect to team composition, we focus on basic observable characteristics, namely, on team size, gender composition, and task-specific experience of team members. With respect to governance structures, we elicit how important experts consider monetary team bonuses, competitions for status (i.e., rank incentives),

[6]Further, correlational studies suggest a positive relationship of task-related diversity on team performance (e.g., functional expertise, education, or organizational tenure) but no significant relationship for bio-demographic diversity, e.g., age or gender (see, e.g., Horwitz and Horwitz, 2007), and that team size can be a mediating factor of the relationship between team diversity and performance. Causal estimates related to diversity in non-routine tasks are however rare. Hoogendoorn et al. (2013) identify a positive causal effect of gender diversity in start up teams. Huber et al. (2020) show that balanced skills can be beneficial for a team's venture performance but only if it comes from within-person skill balance (i.e., combining team members with different skills in mixed teams does not compensate for a lack of members who individually possess balanced cognitive skills). Hoogendoorn et al. (2017) find an inverse U-shaped relationship between cognitive ability dispersion in start up teams and team performance. Dutcher and Rodet (2022) identify a positive causal effect of diversity among team members' experience and knowledge but not for diversity over observable characteristics in a divergent thinking (alternative uses) task (Torrance, 1966). For a more comprehensive review of the literature on diversity see also Harrison and Klein (2007).

[7]Evidence from related literature on creativity (e.g. Bradler et al., 2014; Charness and Grieco, 2019; Gibbs et al., 2017; Laske and Schroeder, 2016; Ramm et al., 2013), indicate no negative effects but mostly positive incentive effects, which have also been identified in other tasks that require mainly cognitive effort such as education and teaching (Fryer et al., 2012; Muralidharan and Sundararaman, 2011).

competitions for monetary prizes, as well as male and female leadership. Importantly, all focus attributes are also observable in the performance data we use to identify experts' misperceptions (which stem from a series of large scale field experiments that identify the causal effects of the above mentioned team governance structures, see also Englmaier et al., 2018, 2023a, 2021).[8]

Regarding team composition, we find that experts expect larger teams to perform better. Further, experts value gender diversity and, on average, prefer perfectly gender-mixed teams the most. Finally, experts prefer teams with at least one experienced team member. In terms of team governance structures, experts expect that teams facing performance incentives in the form of bonuses, rank, and tournament incentives perform substantially better than teams without such incentives, and experts also attribute positive value to having a team leader (of either gender).

Contrasting experts' perceptions to actual performance data of 1,062 teams in the same task reveals that experts form by and large reasonable expectations. In about 75 percent of team comparisons, experts choose the team that is objectively predicted to perform better. Qualitatively, experts do not misperceive the relative importance of additional team members, gender diversity, and experience. Quantitatively, experts slightly underestimate the importance of experience and team size, and tend to overestimate the value of a perfect gender mix. Regarding team governance structures, experts tend to underestimate the performance enhancing effect of team bonuses and substantially underestimate the positive value of leadership, particularly of female leadership.

To study whether perceptions regarding the importance of leadership are prone to the particular task for which we have performance data, we confront experts with an additional non-routine analytical team task in a very different setting. Doing so, we find that experts evaluate the relative importance of attributes regarding team composition and governance structures largely similar across the two non-routine team tasks they were confronted with. Hence, perceptions regarding the importance of leadership are

[8]Given the restrictions concerning potential choice overload within the discrete choice experiment, we did not include additional interesting attributes related to team composition or governance structures. Future research may for instance investigate how experts evaluate the importance of "deep level diversity", which is often conceptualized as differences among members' attributes that are not readily observable but potentially learned trough interaction (e.g., members creativity, personality traits, values or attitudes) (Harrison et al., 1998, 2002), or investigate the perceived role of psychological safety (Castro et al., 2022).

not strongly affected by characteristics of one particular task we used to identify misperceptions.[9]

To investigate why experts undervalue leadership, we evaluate whether HR expertise and own leadership experience can mitigate misperceptions regarding the value of leadership. We compare experts' perceptions to those of a general population sample (n=3,000) and additionally study how own leadership experience (among HR experts) affects perceptions. We find that the undervaluation of the efficacy of leadership also prevails in the general population sample, and that having been employed in a leading position does not increase the perceived value of leadership.

While leadership is undervalued overall, we also detect gender-specific differences in the perceptions of leadership efficacy. Male experts evaluate male leadership substantially more positively than female experts (in both non-routine tasks). Hence, we identify an important implicit bias based on gender.[10] To study (mis)perceived social norms as a possible cause, we further elicit HR experts' second-order beliefs following the idea of the elicitation procedure introduced by Krupka and Weber (2013). That is, we incentivize experts to bet on the team the majority of their peers will choose. The analyses of second-order beliefs reveal only few misperceptions regarding social norms. Experts first and second-order beliefs about the relative importance of different attributes regarding team governance are overall closely aligned, and differences for team composition are small in magnitude. However, we do find that second-order beliefs tend to be more optimistic about leadership efficacy. Further, gender-specific biases in the evaluation of leadership efficacy are less pronounced in second-order beliefs, indicating that experts believe others to judge (female) leadership as more important than themselves.

Overall, our study provides three major contributions. First, we are the first to systematically identify relevant decision makers' perceptions about the relative importance of team characteristics and team governance structures for team success. Second, using data from previous field experiments on team performance in non-routine tasks, we identify systematic underappreciation of leadership efficacy. Third, we document that experts are prone to implicit (gender) biases that, if acted upon, may hamper team productivity.

[9]This holds also true in a between subjects comparison of experts who have (until that point) only evaluated one of the two tasks.

[10]While we can only capture an implicit bias due to the nature of the discrete choice elicitation, we cannot rule out that the bias manifestation is also explicit, as we did not elicit direct explicit bias measures from participants.

By doing so, we contribute to the literature on diversity in teams, biases in performance expectations, as well as to the literature on optimal governance structures for team performance in complex, non-routine tasks. Most closely, our study relates to a nascent literature on (biased) performance expectations. Motivated by gender segregation that characterizes many labor markets around the world (Blau and Kahn, 2017), this literature so far mainly focuses on perceived differences in performance due to gender and gender diversity in teams. In recent work, Fischbacher et al. (2022) study perceived job-specific productivity differences between men and women in tasks that differ with respect to gender stereotypes. In their experiment, participants bet on the success of a team that receives a new team member which is either female or male. They find that participants chose new female members more often for the stereotypically female task and new male members more often for the stereotypically male task. Further, participants tend to bet on gender diverse teams, especially in a task with gender complementaries. In contrast to their work, our approach focuses on non-routine complex problems that are not particularly gender-stereotypical. Further, instead of studying gender composition in isolation, we focus on several potentially important attributes related to team performance (i.e., attributes related to team composition more generally but also team governance structures). We substantially advance this literature by contrasting performance expectations with actual performance data and thereby highlight whether experts actually misperceive the empirical relevance of key determinants of team success in non-routine analytical team tasks.

Secondly, we complement the literature on implicit biases decision makers may hold (e.g., based on their personal characteristics and experiences) and misperceived social norms (i.e. biased second-order beliefs). Many behavioral studies have shown that people are prone to hold biased beliefs based on gender stereotypes (Spencer et al., 1999), and that such beliefs can influence the gender gap in performance. Gender-stereotypical beliefs may also influence ability beliefs (Bordalo et al., 2019) and performance, and implicit biases may result in substantially worse outcomes. For example, Carlana (2019) provides evidence that the gender gap in math performance substantially increases when students are assigned to math teachers with stronger gender stereotypes. That is, teacher implicit bias induces girls to underperform in math and self-select into less demanding high schools (following the track recommendation of their teachers). Recent work by Dustan et al. (2022) further studies first and second-order beliefs to understand whether

biased perceptions about performance exist, which may eventually result in gender gaps in employment. They find no evidence that men's and women's first-order performance beliefs differ between a math task and a bargaining task, but both men and women believe that such belief differences exist. Similarly, there is a growing literature on misperceived social norms highlighting that second-order beliefs may crucially affect final outcomes (Bursztyn et al., 2020).[11] Our work advances this literature studying biases in expectations among HR experts who evaluate the relative importance of team composition and governance structures. Misperceptions by these experts may lead to suboptimal team compositions and may prevent them from cost-effectively fostering team performance. While we do not find strong evidence for misperceptions due to misperceived social norms, we do identify implicit biases related to experts' personal characteristics, as HR experts' own gender affects substantially how they evaluate the value of male leadership.

More broadly, we also connect to work on the role of expectations in work environments and labor markets, and how these expectations may shape selection into different work environments (see also Jäger et al., 2022). For example, Boss et al. (2021) highlight that self-selection of members to entrepreneurial teams may result in suboptimal performance and Gómez-Zará et al. (2020) show (in the context of virtual teams) that information on the diversity of team members may reduce diversity of self-composed teams. We complement this literature by showing that experts hold positive expectations about gender diversity in non-routine team tasks but also find that experts believe others to care less about gender diversity (see also Fischbacher et al., 2022).

Finally, our results point out important implications for practitioners. First, and reassuringly, we find that HR experts are generally aware of essential drivers of team success, as they consider, both, attributes related to team composition and team governance structures, key. Second, HR experts substantially misperceive the value of leadership and may thus forgo cost-effective solutions to foster team performance in non-routine tasks. Third, we find that experts' gender affects perceptions regarding gender-specific leadership efficacy, rendering implicit biases important for the choice of leaders. Fourth, we do not find strong misperceptions regarding social norms among HR experts. Instead, we observe that these can even reduce gender-based biases. Thereby, our results highlight a poten-

[11] Various additional studies have collected important additional evidence for misperceptions that may cause belief based discrimination and misallocations (see, e.g., Bohren et al., 2019a,b; Erkal et al., 2021; Barron et al., 2022; Flynn et al., 2017; Heursen et al., 2020).

tial strategy to reduce own biases by reflecting on how others evaluate determinants of team success.

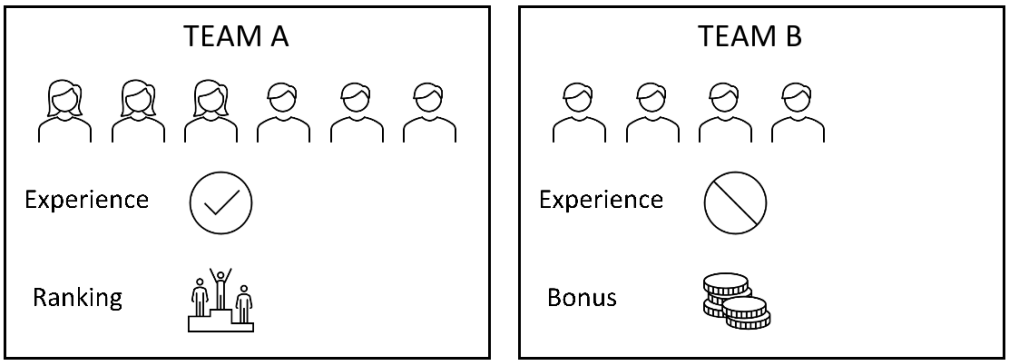
The rest of this paper is structured as follows. Section 3.2 describes our experimental design, measurements, and procedures in more detail. In Section 3.3 we provide the main results related to perceptions and misperceptions of HR experts. Section 3.4 discusses potential reasons for observed misperceptions. Section 3.5 concludes.

3.2 Experimental design

To elicit perceptions about the relative importance of different attributes concerning team composition and team governance structures in non-routine, analytical team tasks, we conduct a large scale incentivized discrete choice experiment. In a discrete choice experiment, participants reveal their preferences over choice attributes by choosing between two options in a series of comparisons. In our context, participants have to choose the better team among two teams in a team comparison (see the example in Figure 3.1), and in each comparison, several team attributes vary. By choosing a team, participants generate two data points per comparison (the chosen option is coded as 1, the other option as 0) which can then be analyzed by a conditional logit model that takes the exact features of the compared options into account. Doing so, the perceived relative importance of individual attributes across choices can be identified (McFadden, 1973; Manski, 2001; Morikawa et al., 2002).

The choice attributes of interest in our setting come from two broad categories: team composition and team governance. The teams we presented to participants differed in their composition and the applied governance structure. The team composition varied in i) *group size* (ranging from four to six team members), ii) *gender ratio* (which we define to be the share of males), and iii) *experience* (a binary indicator of whether at least one individual from the group had previous experience in solving the specific non-routine task the team was facing). For team governance structures, five different conditions were used: *bonus, ranking, prize, male / female leadership*, and *control*. The *bonus* and *prize* conditions used monetary incentives as a governance structure. Under the *bonus* condition teams could earn an additional bonus (of 50 €) when completing the task within a shorter time frame. The *prize* condition was a tournament with a monetary prize, in which team performances were ranked on an online platform accessible by all compet-

Figure 3.1: Discrete choice between two teams



Notes: The figure shows an example of a potential comparison. The two teams differ in their composition and the applied governance structure. While Team A consists of six team members (3 female, 3 male), at least one group member is experienced and they participate in a tournament with a ranking, Team B consists of 4 team members (all male), no one has experience and they are incentivized with a bonus for a fast solution.

ing teams, with the best team winning 150 €. Meanwhile, the *ranking* condition featured rank and status incentives, i.e., a public online ranking without a monetary prize. In the *male / female leadership* condition, teams choose one team member (which could be either male or female) as their leader to guide the group throughout the task. *Control* exhibited no additional governance structure. Overall, this yields 204 possible combinations of attributes which results in 20,706 potential comparisons.

To maximize efficiency for our discrete choice experiment, we built on work by Hole (2017) and Carlsson and Martinsson (2003) who propose employing a modified Fedorov algorithm that maximizes *D-efficiency* in a conditional logit model.[12] The algorithm delivered a set of 180 comparisons which were partitioned in 15 blocks of 12 comparisons each.

We incentivized choices following the method first proposed by Bardsley (2000). Participants were instructed that they could receive a bonus payment of 100 € for one of the team comparisons they were confronted with if they chose the better performing team in that comparison.[13] As participants were not informed which comparison was payoff relevant, they had an incentive to always bet on the team they expected to perform best. To be able to incentivize participants' choices, we added one team comparison for which

[12] A model is D-efficient if it has the "lowest" covariance matrix of the parameter estimates. The modified Fedorov algorithm determines the lowest covariance matrix by iteratively modifying a specific set of comparisons until no further improvements can be made.

[13] We selected 1 in 100 participants to be eligible for payment, which we also made common knowledge.

we observed actual performance outcomes.[14] Hence, we ended up with 15 blocks of 13 comparisons.

3.2.1 Treatments and procedures

We implemented the discrete choice experiment in two different contexts: a *real-life escape challenge*, in which teams solve a series of complex problems in a given amount of time and a *web development* task, in which teams of developers create a professional solution for an innovative web presence for a business customer before a specific deadline. Both tasks were described as non-routine analytical (i.e., teams need to solve a diverse set of complex subtasks in order to succeed) and of closed form (i.e., these tasks entail a clear solution and need to be completed within a specified time frame). We chose the escape challenge as a non-routine analytical team task, as we have performance data for more than 1,000 teams available from a series of natural field experiments (Englmaier et al., 2018, 2023a, 2021), and introduced the web development task as a relevant business setting to assess the generalizability of the elicited perceptions.

Participants of the discrete choice experiment had to bet on teams in both contexts and were randomly assigned to one of two treatment arms which varied whether they first encountered a block of 13 comparisons of teams performing the escape challenge or the web development task.[15] To also elicit potentially misperceived social norms and following the idea of the norm elicitation procedure introduced by Krupka and Weber (2013), participants were further asked to bet on the modal choice of their peers in a third block (see Figure 3.2).[16]

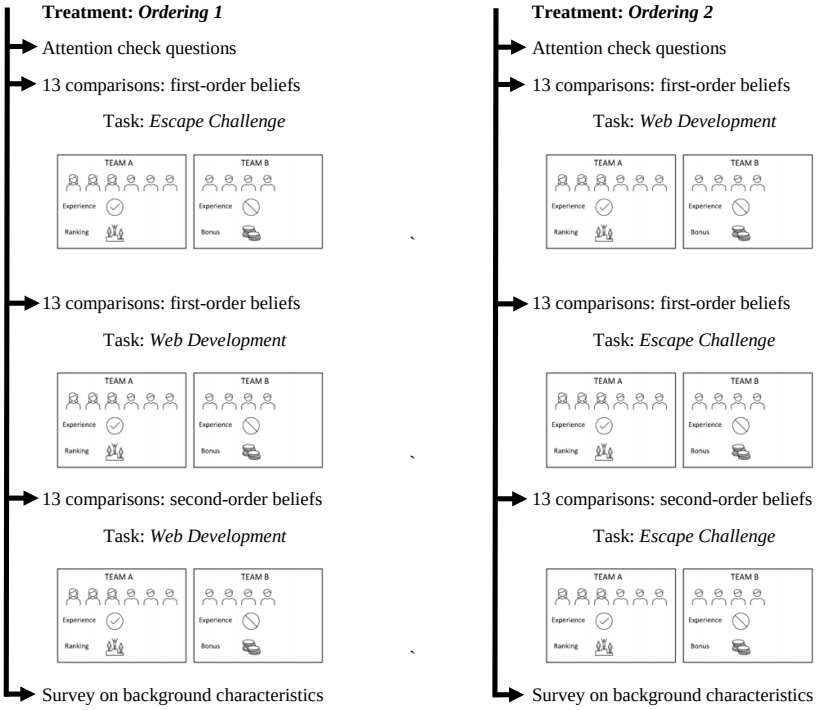
When starting a block, participants were informed about the nature of the task. For the *escape challenge*, we informed participants that teams in escape challenges need to solve a series of complex tasks to escape from a room, that these teams need to find various clues, combine information, and think outside the box. In the *web developer* condition, participants were told to assess the performance of developer teams whose

[14] This team comparison was indistinguishable from the other 12 comparisons in a given set, and we randomly assigned 15 such (randomly chosen) comparisons from our performance data to the 15 different sets.

[15] Half of all participants saw *Ordering 1* whereas the other half saw *Ordering 2* (see also Figure 3.2).

[16] The third block was introduced as an additional chance to earn money, for which we randomly selected 1 in 200 participants to be eligible for payment. These participants could earn 100 € if they chose the team which was chosen by the majority of their peers in one randomly selected team comparison. Doing so, we keep the expected earnings from a correct bet identical for each of the three blocks.

Figure 3.2: Ordering of treatments



Notes: The figure shows the order in which participants were exposed to the treatment arms. In *Ordering 1*, shown on the left, participants first complete 13 comparisons on first-order beliefs for Escape Challenges, and then do the same for the Web Development task. For the latter, we then also elicit second-order beliefs. In *Ordering 2*, shown on the right, participants first complete 13 comparisons on first-order beliefs for the Web Development task, and then do the same for the Escape Challenge. For the latter, we then also elicit second-order beliefs. Half of the participants were exposed to *Ordering 1* and half to *Ordering 2*.

task was to develop a professional solution for an innovative web presence. Participants knew that this web solution had to meet a number of specified requirements and had to be completed within a specified time frame. Hence, both contexts reflect important aspects of non-routine, analytical team tasks. After participants completed all three blocks, they answered a short questionnaire related to their personal characteristics, work, and leadership experiences.[17]

[17]This questionnaire included information regarding their age, highest level of education, experience with escape games, experience with web development tasks, information regarding whether their own work tasks are non-routine / analytical, whether they work in a team, whether they have leadership experience, and how they think about a leader they (have) work(ed) with (i.e., whether this leader encourages them to succeed, leads their team effectively, whether they were satisfied with that leader), and whether they have been involved in selecting a leader in the past. Finally, the questionnaire included their expectations regarding the length of escape challenges and web development tasks, income of people performing these tasks, and expected costs for an escape challenge.

3.2.2 Sample characteristics

We administered the incentivized discrete choice experiment online in mid-December 2022. In total, we recruited 6,000 participants from the subject pool maintained by professional survey provider *Cint*. Our main focus lies on a sample of 3,000 HR experts (see Table 3.1, top panel), who had been actively involved in making HR decisions as part of their daily work. To evaluate whether HR expertise can mitigate misperceptions regarding the value of leadership, we additionally recruited a representative sample of the working German population (based on age, gender, and state of residence, $n = 3,000$, see Table 3.1, bottom panel).

As Table 3.1 shows, our sample of HR experts tends to be younger and more educated than the general population. They are also more often working full time and have more experience with web development and escape games. With respect to their job characteristics, HR experts are more likely to work on tasks described as non-routine (84% vs 75% answered 'rather yes' or 'yes'), analytical (72% vs 54%), team-based (85% vs 77%), and they are much more likely to hold a leadership position: 83% vs 47%.

In both samples, each respondent betted on one of two teams in a total of 39 team comparisons (i.e., we elicited first-order beliefs of each participant for 13 comparisons regarding teams performing an escape challenge and 13 comparisons regarding teams performing the web development task, and second-order beliefs for 13 comparisons for the task they encountered second; see Figure 3.2). This yields a total of 234,000 decisions.

Table 3.1: Background characteristics

	HR experts (n = 3,000)			
Gender	Female	Male	Other	
	50.83%	48.60%	0.57%	
Age	18 - 25	26 - 35	36 - 50	50+
	23.73%	57.57%	29.00%	13.43%
Education	University	High school	Other	No degree
	40.60%	41.37%	17.53%	0.50%
Experience	Escape Game	Web Design		
	49.90%	36.37%		
Employment status	Full-time	Part-time		
	84.60%	15.40%		
Job characteristics	No	Rather no	Rather yes	Yes
Non-routine	3.87%	12.17%	48.73%	35.23%
Analytical	4.50%	23.60%	50.23%	21.67%
Team-based	4.53%	10.73%	34.10%	50.63%
Leadership role	4.30%	13.03%	37.57%	45.10%
	General population (n = 3,000)			
Gender	Female	Male	Other	
	50.80%	49.07%	0.13%	
Age	18 - 25	26 - 35	36 - 50	50+
	12.77%	20.40%	32.67%	34.17%
Education	University	High school	Other	No degree
	32.03%	34.13%	33.07%	0.77%
Experience	Escape Game	Web Design		
	36.40%	23.20%		
Employment status	Full-time	Part-time		
	71.63%	28.37%		
Job characteristics	No	Rather no	Rather yes	Yes
Non-routine	4.57%	20.47%	49.30%	25.67%
Analytical	12.70%	33.17%	41.33%	12.80%
Team-based	9.27%	13.37%	34.57%	42.80%
Leadership role	26.03%	26.63%	31.77%	15.57%

Notes: The table displays sample characteristics with respect to gender, age, educational background, experience with both tasks, the employment status and job characteristics.

3.3 Results

We structure our results as follows. First, we present HR experts' perceptions regarding the determinants of team success in the non-routine analytical team task (team escape challenge) that we subsequently contrast with predictions based on actual performance data of 1,062 teams to highlight experts' misperceptions. Second, we discuss whether experts' perceptions about the determinants of team success substantially differ in the other non-routine analytical team task they were asked to evaluate (web developer task).

3.3.1 Perceptions

To establish HR experts' perceptions about the relative importance of different choice attributes, we run a (fixed effects) conditional logit model that regresses the choices of the respondents on all dimensions of team composition (group size, gender diversity, and experience) and team governance structure (bonus, rank, prize, female and male leadership) for the two distinct tasks.[18] We cluster standard errors at the comparison level.

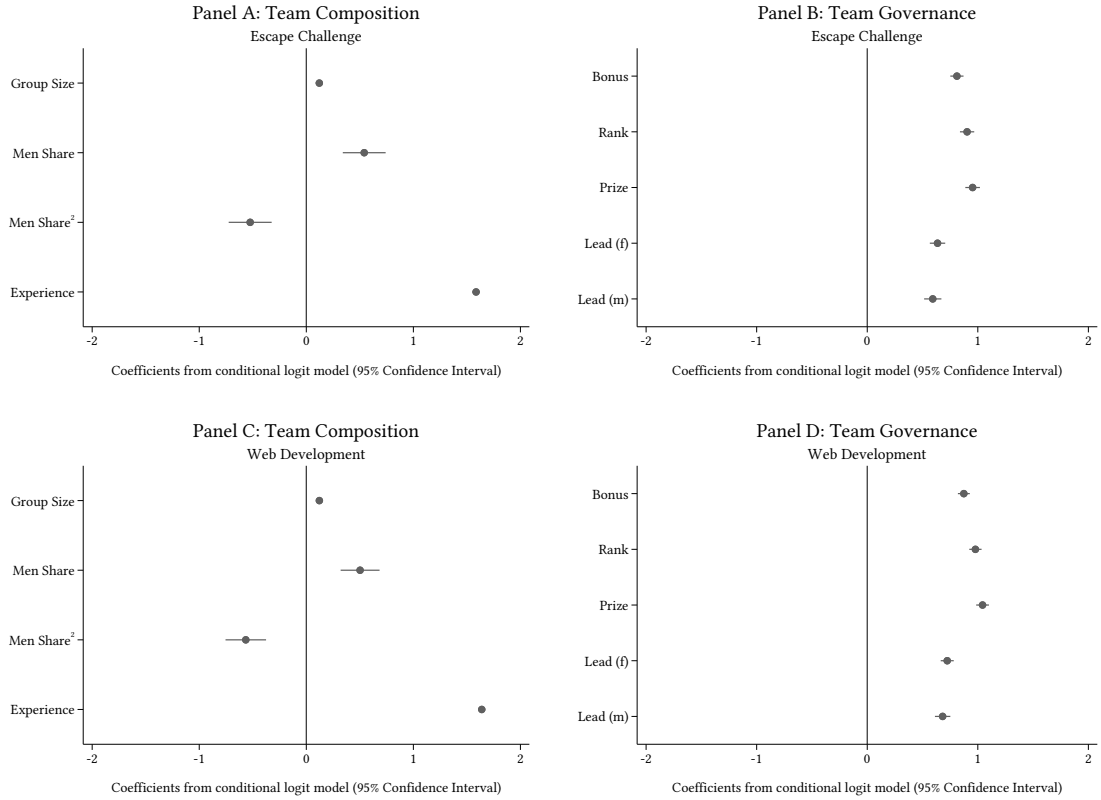
Figure 3.3 plots the estimated coefficients (including 95% confidence intervals).[19] Panel A illustrates HR experts' perceived relative importance of all team composition dimensions for escape challenges. As becomes clear, experts expect a positive effect of larger teams and value gender diversity (evidenced by a positive effect for the male share and a negative effect of similar magnitude for the squared male share).[20] Further, experts expect that having at least one experienced member on a team substantially improves team success. Holding everything else constant, the odds of betting on a team are 4.88:1 when an experienced member is on that team. Panel B illustrates experts' perceptions regarding the relative importance of various team governance structures for escape challenges. Experts expect that teams facing performance incentives (in the form of bonuses, rank, and tournament incentives) substantially improve performance and attribute positive value to having a team leader (independent of the leader's gender). Holding everything else equal, the odds that experts bet on a team facing bonus incentives as compared

[18] We capture gender diversity by including a linear and a quadratic term for the male share in a team.

[19] Table 3.A.1 in the Appendix reports the coefficients presented in Figure 3.3 in specifications (1) and (5). The table additionally includes specifications taking interactions between team composition and governance structures into account, which are discussed briefly in Appendix Section 3.A.1.

[20] According to the obtained coefficients, on average their most preferred gender ratio is 51% male.

Figure 3.3: Perceptions about team composition and team governance



Notes: The figure shows coefficient plots from conditional logit models (95% confidence bands). Panel A shows perceptions about team composition in Escape Challenges. Panel B shows perceptions about team governance structures in Escape Challenges. Panel C shows perceptions about team composition in the Web Developing task. Panel D shows perceptions about team governance structures in the Web Developing task.

to betting on a team without incentives are 2.25:1. For rank incentives, the odds are 2.46:1, for tournaments with monetary prizes 2.59:1 and for teams with female (male) leaders the odds are 1.89:1 (1.81:1). These results imply that experts place a similar emphasis on the three incentive structures, but expect a much lower impact of leadership (independently of the leaders' gender). Panel C and Panel D illustrate the relative importance of team composition and team governance structures for the web developing task. We relegate the comparison between perceptions in the two non-routine analytical team tasks to Section 3.3.3.

3.3.2 Misperceptions

To juxtapose the *perceived* determinants with *actual* determinants of team success, we resort to performance data from three field experiments which analyzed the impact of bonus incentives (Englmaier et al., 2018), tournament incentives (Englmaier et al., 2023a) and leadership (Englmaier et al., 2021) on performance in a non-routine team task (a real-life escape challenge). These studies were conducted in cooperation with *ExitTheRoom*, an escape challenge provider in Munich (Germany). Of the 1,358 teams from all three studies, we eliminate those with less than four or more than six individuals to arrive at $n = 1,062$ teams to create an objective benchmark for experts' perceptions.[21]

Based on the performance data of these (actual) teams, we use a Tobit model to predict the time in which each team (displayed in the discrete choice experiment) is expected to complete the escape challenge (our measure of team success). The Tobit model takes the upper limit of 60 minutes (the deadline for all teams in the escape challenge) into account and includes all team composition and team governance attributes shown in the discrete choice experiment as explanatory variables.[22] The resulting coefficients represent the 'true' effect of each of these dimensions. We then simulate choice data of a 'Naïve Expert' who picks the team with the lower predicted finishing time in each team comparison of the discrete choice experiment and compare the 'choices' of this 'Naïve Expert' with those of the HR experts.[23] We then run a conditional logit model in which we estimate whether we observe systematic differences in the choices of the 'Naïve Expert' and the HR experts. Doing so, we are able to unveil misperceptions of HR experts relative to predictions based on the actual performance data.

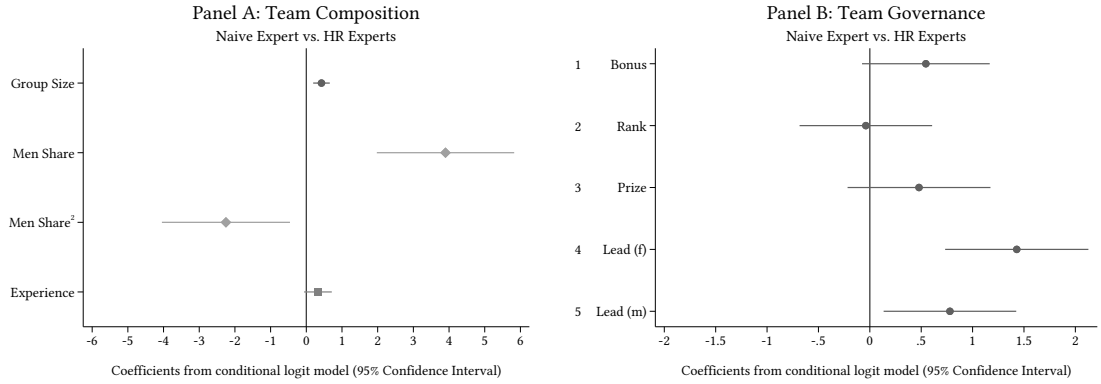
Overall, we find that experts form reasonable expectations. On average, HR experts choose the team that the 'Naïve Expert' would choose in 75% of all comparisons. However, for team comparisons, for which HR experts' choices did not coincide with those of the 'Naïve Expert', the difference in predicted finishing times between the chosen and the unchosen team amounts to 46% of the average remaining time before the 60 minutes deadline in the escape challenge, meaning that these misperceptions could potentially be

[21] This was done to ensure comparability as comparisons in the discrete choice experiment only featured team sizes four to six.

[22] As the data originates from three independent studies that were conducted at different points in time, it also includes study fixed effects. The resulting coefficients are illustrated in Appendix Figure 3.A.1.

[23] We call the expert 'Naïve', as the predictions from the Tobit model do not take interactions between choice attributes into account.

Figure 3.4: Differences between 'Naïve Expert' and HR experts



Notes: The figure shows coefficient plots from conditional logit models (with 95 % confidence bands). Panel A shows differences in perceptions with respect to team composition. Panel B shows differences in perceptions with respect to team governance structures. A positive (negative) value indicates that a 'Naïve Expert' expects this factor to be more (less) important than the HR experts.

very costly. Finally, we find that misperceptions are not due to random choice. Instead, errors are systematic as they become more frequent the closer the differences in predicted completion times between teams are (see also Appendix Figure 3.A.2).

Figure 3.4 compares the relative importance the 'Naïve Expert' assigns to the different team attributes as compared to the HR experts and thereby illustrates whether the latter misperceive the relative importance of an attribute based on the predictions from the actual performance data. Panel A illustrates the comparison of the relative importance of the team composition attributes (positive coefficients indicate that the respective attribute is empirically more important than HR experts expect). As can be seen, the 'Naïve Expert' choices indicate that team size and experience are slightly more relevant than HR experts think, but these misperceptions are small. Regarding team composition, experts correctly anticipate that diversity is valuable, but they prefer gender equality (on average 51 percent males), whereas the estimates from our performance data yield a higher coefficient for the share of males in a team and a lower coefficient for the quadratic term. Hence, 'Naïve Expert' choices indicate that diverse teams with more males performed even better (the optimum lies at on average about 80% men in a team).

Panel B illustrates the comparison of the relative importance of team governance structures for the 'Naïve Expert' and the HR experts. While experts form the right qualitative expectations regarding team governance structures, their perceptions are not always quantitatively accurate. Experts systematically underestimate the performance-

enhancing effect of team bonuses ($p = 0.085$). Regarding rank incentives, HR experts form on average accurate beliefs. It appears as if HR experts slightly underestimate the efficacy of tournaments with monetary prizes, but these differences are statistically insignificant ($p = 0.177$). Most strikingly, experts substantially underestimate the positive value of leadership ($p < 0.001$ for female leadership and $p = 0.018$ for male leadership; for pairwise comparisons, see also Appendix Table 3.A.2). That is, while HR experts' odds for choosing an alternative with a female (male) leader are 1.89:1 (1.81:1), our performance data indicate that these odds should be as high as 7.9:1 for females (and 3.95:1 for males).

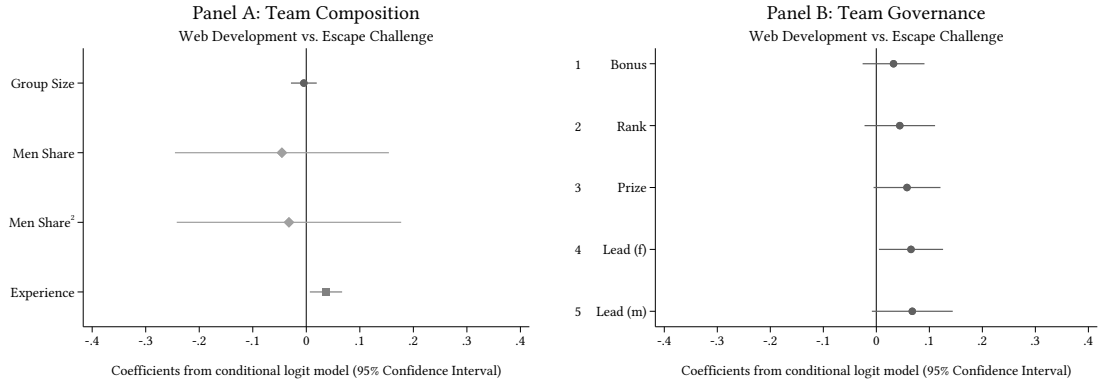
3.3.3 Perceptions across different non-routine analytical tasks

Based on the performance data from 1,062 teams performing an escape challenge, we found that HR experts substantially underestimate the value of leadership for team performance. While team escape challenges bear the defining features of non-routine analytical tasks, HR experts may form perceptions about the determinants of team success that strongly depend on particularities of the escape room setting.[24] It is thus crucial to understand whether HR experts' perceptions are task specific. To study whether our findings generalize to other non-routine tasks in professional contexts, we also elicited HR experts' perceptions for a second non-routine analytical team task: the web development task.

Panels C and D in Figure 3.3 already suggest that the expected relative impact of various team composition attributes and team governance structures closely resemble the results observed for the team escape challenge (shown in Figure 3.3, Panels A and B). In Figure 3.5, we additionally plot the coefficients from regressions of the differences in perceptions between both tasks explicitly. A positive value indicates that HR experts attribute more importance to the factor in the web development task as compared to the escape challenge (and vice versa). Panel A illustrates the relative importance of various team composition dimensions between both tasks. As becomes clear, experts do not expect that the effects of team size or diversity strongly differ across both tasks. However, they expect that experience matters even more in the web development task. Regarding the relative importance of team governance structures for escape challenges and the web

[24] See also Englmaier et al. (2018) for an extensive discussion of the advantages and disadvantages of using escape challenges to study teamwork in non-routine team tasks.

Figure 3.5: Comparison between escape challenge and web development



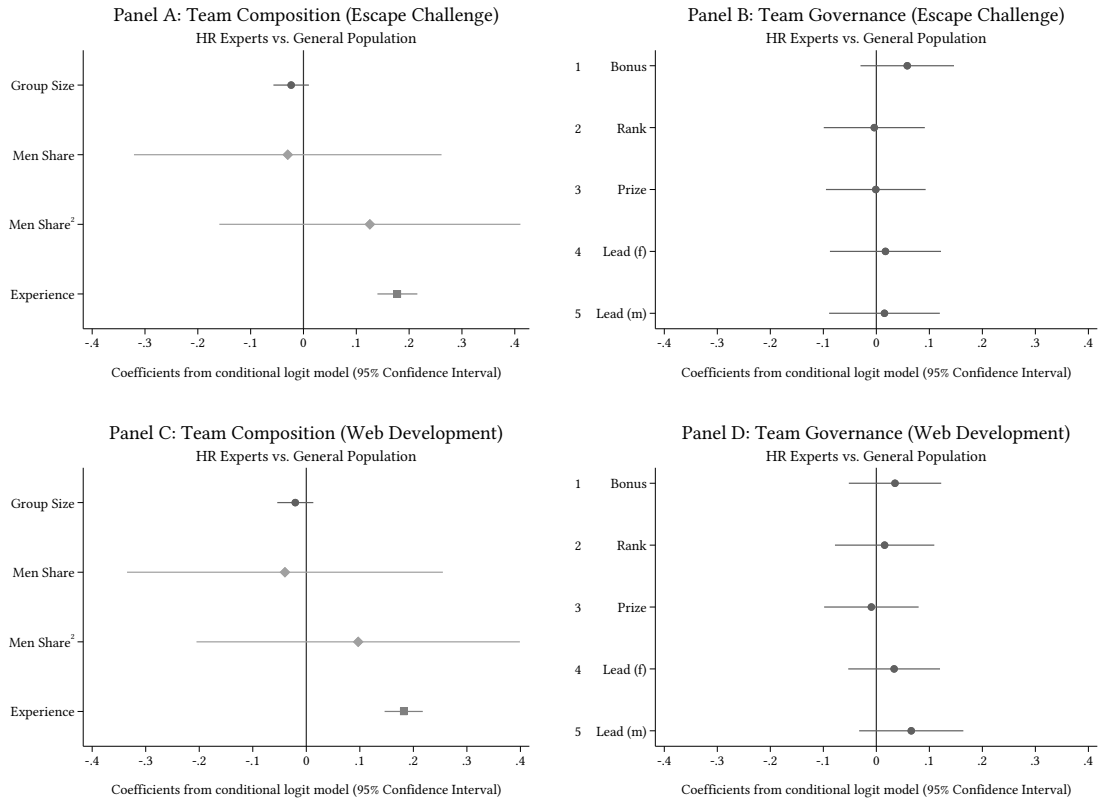
Notes: The figure shows coefficient plots from conditional logit models (with 95% confidence bands). Panel A shows differences in perceptions with respect to team composition. Panel B shows differences in perceptions with respect to team governance structures. A positive (negative) value indicates that participants expect this factor to be more (less) important in the Web Developing task than in the Escape Challenge.

development task, experts expect that teams facing performance incentives (in the form of bonuses, rank, and tournament incentives) substantially improve performance in both tasks, and also attribute positive value to having a team leader (see Panels C and D in Figure 3.3). In comparison, they assume that governance structures matter slightly more in the web development task (see Panel B in Figure 3.5). Overall, these findings emphasize a very similar assessment between both tasks, ameliorating potential concerns that HR experts form perceptions that depend on particularities of the escape challenge setting.

3.4 Discussion

As shown in Section 3.3.2, experts substantially misperceive the value of leadership for team success in non-routine tasks. But where do these misperceptions come from? In this section, we demonstrate that having HR expertise does not mitigate the observed effects, and neither does own leadership experience mediate them. We also show that HR experts' gender relates to potential misperceptions of female versus male leadership. Finally, we study whether misperceptions are larger (or smaller) in second-order beliefs (i.e., whether social norms could be responsible for any observed biases).

Figure 3.6: Comparison between HR experts and general population



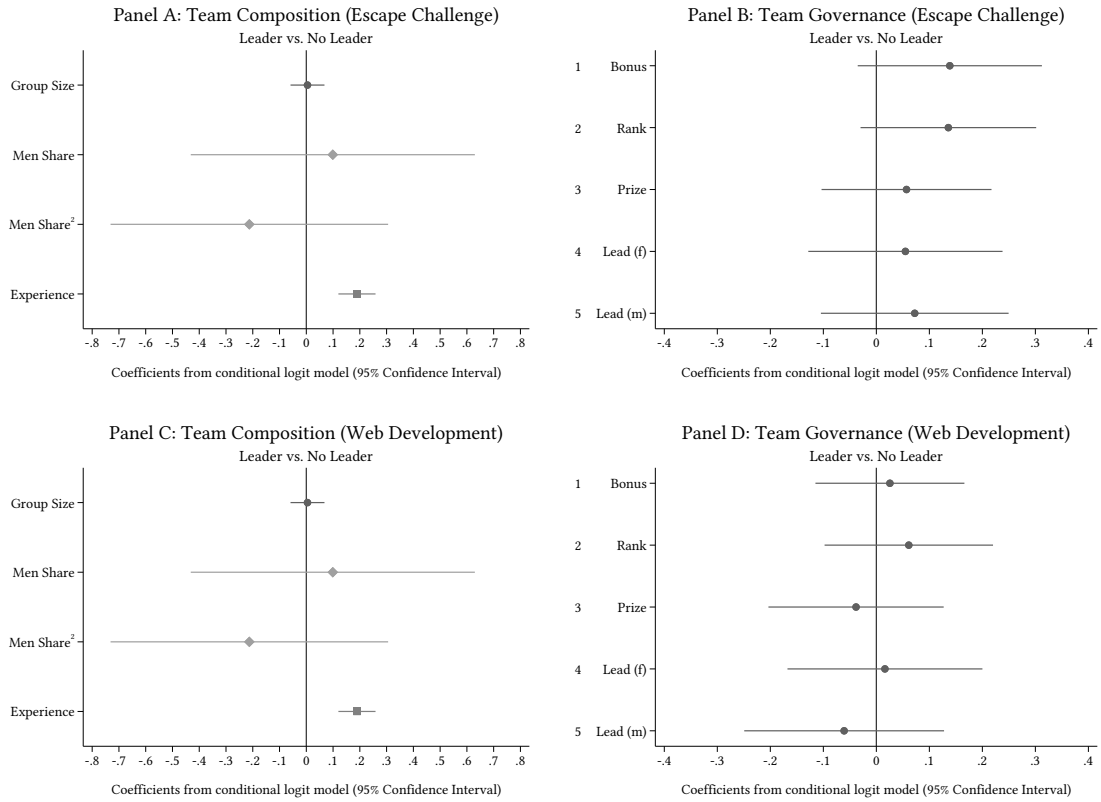
Notes: The figure shows coefficient plots from conditional logit models (with 95% confidence bands). Panel A and Panel C show differences in perceptions with respect to team composition for the Escape Challenge and the Web Developing task. Panel B and Panel D show differences in perceptions with respect to team governance structures. A positive (negative) value indicates that HR experts expect this factor to be more (less) important than the general population sample.

3.4.1 HR expertise and leadership experience

HR experts potentially differ from the general population, and, especially smaller firms, may not use HR experts to assemble work teams or design their governance structure. Therefore, we analyze differences in the relative importance of these factors between our sample of HR experts and a representative sample of the adult German population. The results are displayed in Figure 3.6. Positive values indicate that HR experts expect a stronger impact than members of the general population (and vice versa).

The figure comprises four panels. Panels in the left column depict the effects of dimensions of team composition and panels in the right column depict the effects of team governance structures. The upper row panels report results for the escape challenge, whereas the lower row panels similarly depict the results for the web development task.

Figure 3.7: Leadership experience in the HR expert sample



Notes: The figure shows coefficient plots from conditional logit models (with 95% confidence bands). Panel A and Panel C show differences in perceptions with respect to team composition for the Escape Challenge and the Web Developing task. Panel B and Panel D show differences in perceptions with respect to team governance structures. A positive (negative) value indicates that those with prior leadership experience expect this factor to be more (less) important than those without (in the HR experts sample).

Panel A and Panel C illustrate that both, HR experts and the general population sample, generally agree in their assessment regarding team size and gender diversity in both tasks. HR experts, however, expect much stronger effects from experience than the general population, and correctly so (see also Section 3.3.2). Interestingly, they do so similarly across both task. Panel B and Panel D reveal no significant differences with respect to team governance structures between both samples in either task. Hence, while HR expertise comes with a more appropriate evaluation of the importance of experience, such expertise does not seem to mitigate the undervaluation of leadership.

Another reason for the existence of misperceptions could be that, without having held a leadership role themselves, some HR experts may simply not have first-hand experience with the positive effects leadership can unfold. Figure 3.7 shows the results for those who

respond with 'rather yes' and 'yes' to the question of whether they hold a leadership role relative to the ones answering 'rather no' and 'no'. Positive coefficients indicate that holding a leadership role leads to assigning a higher value to an attribute (and vice versa). Panels A and C show the effects of various team composition dimensions in the escape challenge and the web development task, respectively. Prior leadership experience significantly increases only the value attached to prior experience with the task. Panels B and D show results for the team governance dimensions. In neither task does leadership exert a strong effect on any of these task features, including leadership. It thus seems unlikely that prior leadership experience would increase the valuation of (male or female) leadership (in either task). The misperception of the importance of leadership seems thus ubiquitous among HR experts and the general population, independent of their own experiences.[25]

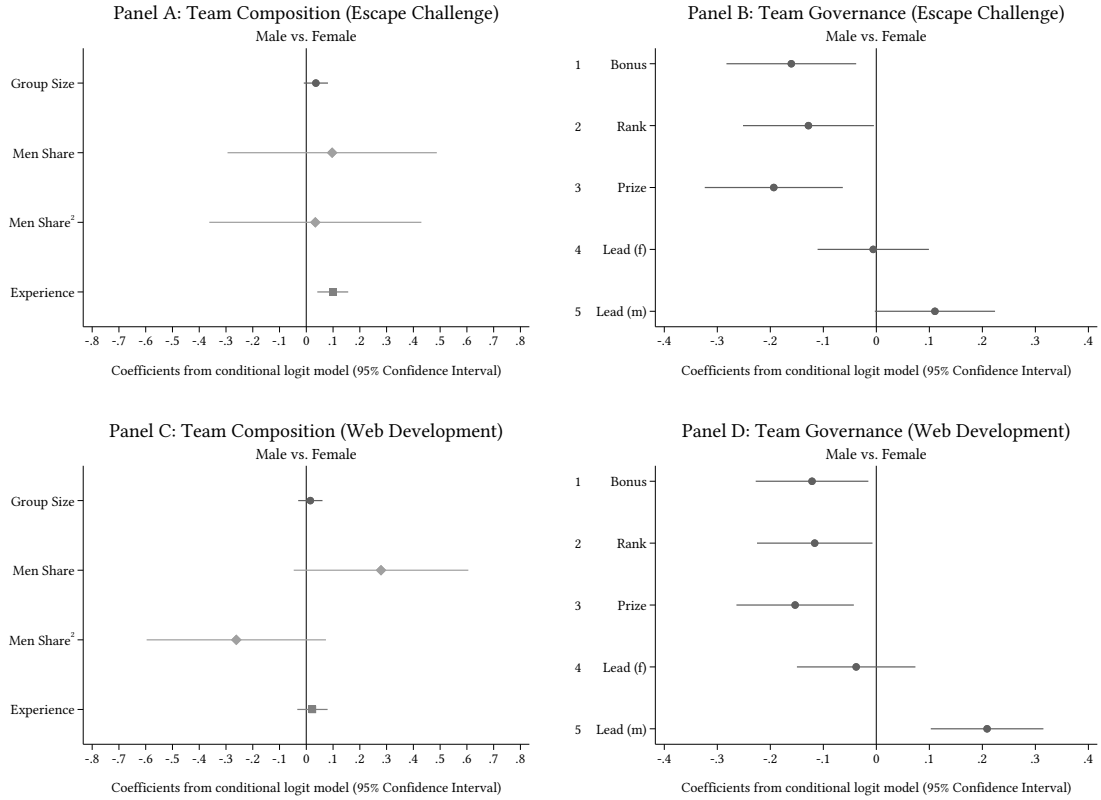
3.4.2 Gender bias

To investigate whether implicit gender biases could be responsible for the stark divergence between the 'Naïve Expert' and our HR experts regarding the role of (especially female) leadership shown in Figure 3.4, we analyze differences in the relative importance of team composition (group size, gender diversity, and experience) and team governance structures (bonus, rank, prize, female leadership and male leadership) between gender. The results are displayed in Figure 3.8. Positive values imply that male respondents place a relatively higher weight on a factor than female respondents and vice versa.

Panel A and Panel C illustrate the comparison of the relative importance of team composition for both genders. Both genders tend to agree on the importance of most factors across both tasks. The only statistically significant difference is that men perceive experience to be more important in escape challenges compared to women. Panel B and Panel D depict the comparison of the relative importance of specific team governance structures for both genders. First, males perceive bonuses, rank and prize incentives to be less effective compared to females. Second, while both genders hold similar beliefs about the effects of female leadership, males (females) anticipate male leadership to be more (less) effective compared to the beliefs of women (men). These results exemplify that

[25]Personal experience with escape challenges does not reduce the misperception of the importance of leadership either. Having experience with developing web designs tends to increase the perceived importance of female leadership for the web development task (see Appendix Figure 3.A.3).

Figure 3.8: Comparison between males and females



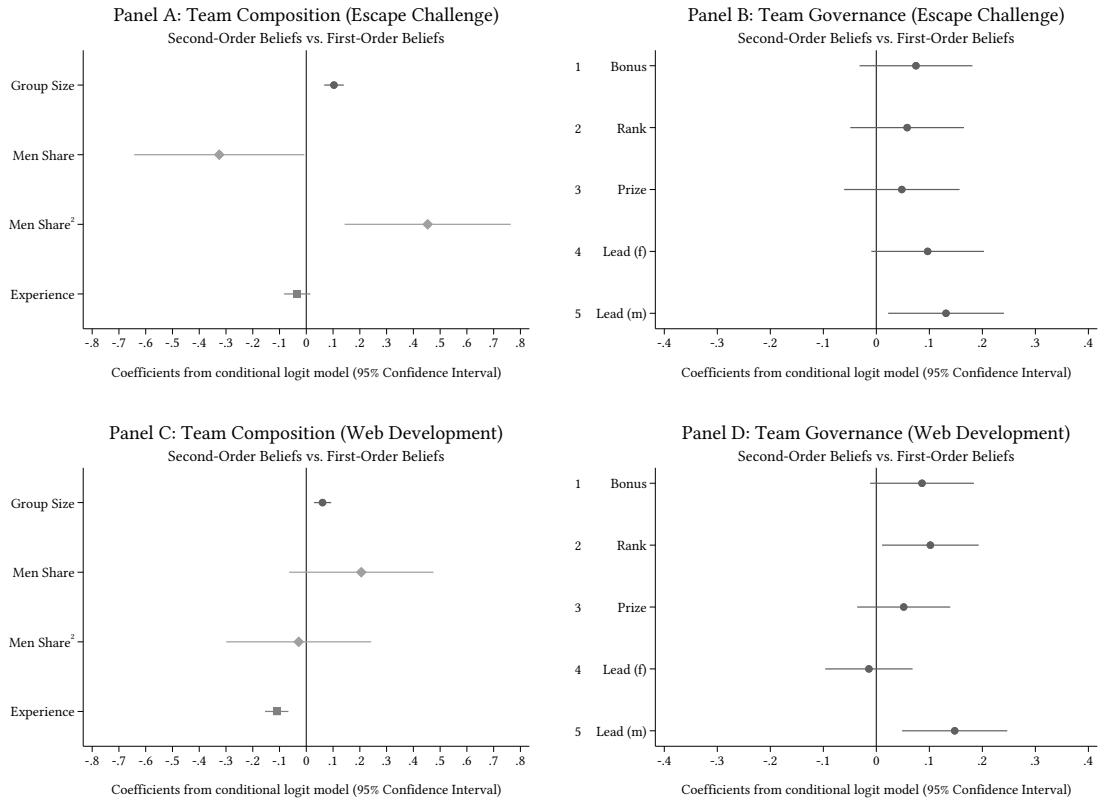
Notes: The figure shows coefficient plots from conditional logit models (with 95% confidence bands). Panel A and Panel C show differences in perceptions with respect to team composition for males and females. Panel B and Panel D show differences in perceptions with respect to team governance structures. A positive (negative) value indicates that males expect this factor to be more (less) important than females.

both men and women seem to have a too negative view of female leadership (compared to its actual impact). The overall more positive view of male leadership (relative to female leadership) thus stems from men valuing male leadership relatively more than women. Overall, these findings suggest that implicit gender biases are not responsible for the misperception regarding the efficacy of leadership. However, own gender does cause an additional bias related to the relative efficacy of female (vs. male) leadership.

3.4.3 Social norms

Another source for misperceptions to arise could be a (mis)perception of the prevailing social norms. If experts believe that others hold female leadership in low esteem, they may similarly adjust their beliefs, for example out of conformity. To study social norms,

Figure 3.9: Comparison between second-order and first-order beliefs



Notes: The figure shows coefficient plots from conditional logit models (with 95% confidence bands). Panel A and Panel C show differences in perceptions with respect to team composition for first-order and second-order beliefs. Panel B and Panel D show differences in perceptions with respect to team governance structures. A positive (negative) value indicates that HR experts expect that other HR experts perceive this factor to be more (less) important than themselves.

we also elicited second-order beliefs. In the following, we analyze differences in the relative importance of the team composition dimensions and team governance structure between first-order and second-order beliefs of HR experts. The results are displayed in Figure 3.9. Positive coefficients indicate that respondents believe others to hold a factor in much higher esteem than they do themselves (and vice versa).

Panel A and Panel C illustrate the comparison of the relative importance of team composition for first-order and second-order beliefs. For escape challenges (Panel A), we find that HR experts' second-order beliefs are less optimistic about gender diversity and more optimistic for the efficacy of larger groups. Respondents thus seem to believe that gender equality is less important to other respondents than to them. For web development, similar findings hold for group size. Furthermore, in this task, HR experts assume that

other experts view experience to be less important than themselves. Panel B and Panel D illustrate the comparison of the relative importance of team governance structures for first-order and second-order beliefs. Generally, first-order and second-order beliefs for the first three determinants seem relatively closely aligned, apart from bonuses and rank incentives in the web development task, where HR experts believe others view them as more important. For escape challenges, however, we find that HR experts' second-order beliefs are more optimistic about leadership efficacy (independent of the leaders' gender). For web development, this only holds for male leadership. Interestingly, we find that, contrary to HR experts' first-order beliefs, experts' own gender does not cause a bias in second-order beliefs about the efficacy of (female) leadership (see Appendix Figure 3.A.4).[26] Overall, these results emphasize that it does not seem to be the case that HR experts suffer from distorted beliefs about the social norm with respect to leadership and then adapt to the norm. Instead, they think that others have a more positive view with respect to the efficacy of leadership than they themselves do.

3.5 Conclusion

As teamwork in non-routine analytical tasks has become increasingly important in recent decades and is now prevalent in many areas of the economy, it is imperative for firms' HR experts to form the right expectations about the determinants of team success. As such knowledge may not be readily available to firms in non-routine environments, experts face a challenging problem when assembling teams and devising teams' governance structures. They may be prone to (implicit) biases, follow (mis)perceived social norms, or substantially underestimate particularly effective determinants of team success. As a consequence, their choices may result in inefficient economic outcomes. This study provides the first comprehensive analysis of HR experts' perceptions regarding the effects of a set of key potential determinants on team success, ranging from aspects about team composition to the governance structures teams face. In a large sample of HR experts (n=3,000), we show that i) experts expect larger teams to perform better, ii) experts value gender diversity and, on average, prefer perfectly gender-balanced teams the most, and iii) experts prefer teams with at least one experienced team member. In terms of

[26]Hence, eliciting second-order beliefs may potentially help to reduce biases stemming from a decision maker's own characteristics.

team governance structures, experts expect that performance incentives (in the form of bonuses, a public ranking, and a prize) enhance team performance and consider them similarly effective. Experts also predict positive effects through (male and female) team leaders, but expect leadership to matter less than performance incentives. HR experts' perceptions turn out to be robust across two non-routine team tasks that exhibit different characteristics, and are qualitatively similar to perceptions elicited from a large general population sample (n=3,000).

We exploit the unique opportunity to contrast experts' perceptions with performance data from 1,062 teams working on the same non-routine task. We find that experts slightly underestimate the importance of experience and team size, and tend to overestimate the value of perfect gender balance in teams. Regarding team governance structures, experts tend to underestimate the performance-enhancing effect of team bonuses, and substantially underestimate the positive value of leadership. This undervaluation of leadership is persistent and prevails also in a general population sample. Further, it is not mitigated by having been employed in a leading position.

We also detect an important implicit gender bias: While leadership is undervalued overall, male experts evaluate male leadership substantially more positively than female experts. This is not driven by (mis)perceived social norms, as first and second-order beliefs of respondents are closely aligned across most determinants. Gender-specific biases of leadership effectiveness are, however, less pronounced in second-order beliefs, indicating that experts believe others to judge (female) leadership as more important than themselves.

This study is the first to provide a comprehensive documentation of HR experts' (and the general population's) perceptions of the effectiveness of various governance structures and dimensions of team composition for team performance. This advance was made possible by combining unique data from a series of field experiments that allow comparing variation in different features of a task environment across identical tasks. These insights provide a promising starting point for future research to explore the role of interventions (e.g., information provision experiments) to determine how perceptions can be altered. Because we find misperceptions regarding some components, these allow for potentially improved work performance through de-biasing or other forms of misperception mitigation.

Having shown that HR experts and the general population exhibit a gender bias with respect to female leadership, our findings may be useful to explain parts of the leaking pipeline, i.e., the phenomenon that we observe fewer women as we move up the corporate ladder. While parts of this phenomenon will be driven by women's preferences (see, e.g., Hampole et al., 2021), we show that HR experts not expecting female leadership to be relatively important enough, may also contribute to the effect. Future work may thus investigate more deeply what role such misperceptions play in designing work environments and to what extent they widen the gender gap in leadership roles compared to more routine types of work.

Finally, our results also carry implications for practitioners. Given the identified misperceptions of HR experts, it seems conceivable that making experts aware of their biases (particularly with respect to the value of (female) leadership) can render teamwork more successful. For instance, we found that asking experts to reflect on how important others perceive (female) leadership indicated that the gender-based relative underweighting of the value of female (vs. male) leadership does not prevail in second-order beliefs. Hence, as long as social norms reflect less biased beliefs, eliciting second-order beliefs may potentially help decision makers to reduce biases based on own personal characteristics. However, given the substantial misperception regarding the value of leadership among HR experts, other interventions that directly aim at the reduction of misperceptions promise substantial economic gains and, therefore, seem worth to pursue.

Chapter 4

Who Does What? Task Assignment and the Role of Productivity and Preferences

Abstract

Task assignment is an important dimension of social comparisons at the workplace and therefore also a key challenge for managers. While there are many studies on the role of wage inequality and wage comparisons, evidence on the effect of task inequality and task comparisons on performance and satisfaction is scarce. Comparisons along this dimension may not only increase performance and satisfaction in preferable tasks, but also decrease performance and satisfaction in less preferable tasks. I study the impact of task assignment and task comparisons on performance and satisfaction by exogenously varying whether two workers, as part of a one-time job, work on the same or different tasks. In the latter case, I further vary between random allocation, allocation according to perceived productivity or preferences and self-assignment among co-workers to analyze the efficacy of different task assignment procedures.

4.1 Introduction

The last decades have been accompanied by significant changes in the working environment. Tasks have become more non-routine, complex and analytical and are now frequently performed in teams (Autor et al., 2003; Autor and Price, 2013). Accordingly, teamwork became a prerequisite for success of modern firms (Bandiera et al., 2013; Weidmann and Deming, 2021) and an important factor for the modern economy in general (Deming, 2017; Driskell et al., 2018).

Along with this development, the allocation of tasks to team members became a common problem faced by managers.[1] While work performance is one important outcome measure, there are also other important dimensions such as work motivation, satisfaction with the work as well as absenteeism and turnover that managers must consider when assigning tasks to team members (Hackman and Oldham, 1976). Therefore, managers should not only take productivity but also preferences for a task into account when trying to improve these outcomes. Another crucial dimension potentially influencing the behavior of workers are task comparisons. Workers might not only care about their own task but also about the task a close co-worker has to execute (Oldham et al., 1982). Accordingly, productivity and preferences for the own task and the task of close co-workers are crucial factors managers should take into consideration when assigning tasks among co-workers. Moreover, all of these trains of thought should be reflected in the assignment procedure, since how tasks are allocated could influence the perceived fairness of the allocation and thus the behavior of workers.

Understanding whether task assignment plays a crucial role for performance and satisfaction has important implications for team organization, but has been understudied in research so far. Thus, the aim of this study is three-fold: First, I investigate whether working on a different task than a co-worker has an impact on performance and satisfaction of a worker. Second, I analyze whether this relationship is driven by the perceived productivity and preference for the own task and the task of a co-worker. And third, I investigate the efficacy of different task assignment procedures.

To answer these questions, I conduct a field experiment, where students are invited to perform one of two distinct and regular tasks for research assistants as part of a one-time job. I exogenously vary whether groups of two co-workers work on the same or different

[1] Task assignment also reflects two important tasks of leaders, namely motivating (House, 1976; Howell and Avolio, 1993; Bass, 1998, 1999) and coordinating (Bass, 1990; House et al., 1999) their team members.

tasks. When working on different tasks, the task assignment procedure is varied between random allocation, allocation according to perceived productivity, allocation according to preferences and self-assignment among co-workers. Thereby, I identify the causal effect of task assignment and task comparisons (i.e. working on a different task than a close co-worker) on performance and satisfaction. I further analyze the impact of perceived productivity and preferences for both tasks on this relationship. Lastly, I investigate the efficacy of different task assignment procedures by comparing overall and individual outcomes between the above described procedures.

To elicit the beliefs of actual decision makers and to get an estimate of potential effect sizes, I conducted a survey with 400 practitioners in the run-up to the experiment. Moving away from a situation in which two co-workers work on the same task, to a situation in which they work on different tasks, they expect strong positive effects on performance and satisfaction for advantaged workers (i.e. working on the preferred task or the task for which they expect to perform better), while expectations regarding the effect on disadvantaged workers (i.e. working on the less preferred task or the task for which they expect to perform worse) are mixed. Furthermore, while practitioners do not expect large differences between different task allocation procedures, they are very differently applied in practice. Overall, the results from this survey reinforce the need for empirical evidence analyzing the effect of task assignment and task comparisons on performance and satisfaction as well as the efficacy of different task assignment procedures.

The findings of this study contribute to the literature on task allocation, job satisfaction and social comparisons at the workplace. First and foremost, this study provides novel insights on important determinants of task allocation. While there is evidence that incentives lead managers to assign workers to incentivized tasks according to productivity (see, e.g. Bandiera et al., 2007; Burgess et al., 2010), Delfgaauw et al. (2020) do not find such task assignment according to productivity despite bonus incentives in place. Instead, they show suggestive evidence that without incentives, managers now assign tasks more according to preferences. Since this is the only statistically significant difference in how tasks are assigned between the treatment and control group and there is no statistically significant performance improvement due to the bonus, preferences for a certain task might be an important driver for performance and job satisfaction.

In line with this argument, this study also contributes to the literature on job satisfaction. The job characteristics model of work motivation states skill variety, task identity,

task significance, autonomy and feedback as core job dimensions (Hackman and Oldham, 1976). Accordingly, preferences for a certain task should reflect by how much a worker is satisfied with these dimensions. In addition, it should be borne in mind that workers do not only care about their own task but also about the task of people they refer to (other people, past-self or future-self). Oldham et al. (1982) provide correlational evidence that people that compare themselves to self-referents (past-self or future-self) are more productive than those that compare themselves to others.

Research on social comparisons at the workplace has mainly focussed on wage inequality and wage comparisons (see, e.g. Card et al., 2012; Cullen and Perez-Truglia, 2022), while causal evidence on the impact of task inequality and task comparisons on performance and satisfaction is scarce. Two exceptions and most closely related to this study are Montagno (1985) and Patchen (1958). Montagno (1985) finds that people who work on an enriched task and were told that other people work on a less enriched task[2], perform better without being more satisfied with their task. While he is only looking at advantaged people, Patchen (1958) also analyzed the behavior of disadvantaged people. Looking at junior high school pupils, he shows that disadvantaged pupils (working on the less preferred task) enjoyed the task the least, but were not less satisfied with the rules. Instead, advantaged pupils were least satisfied with the rules, indicating that behavior is not only driven by the task a person is assigned to, but also by the legitimation of the assignment. This in turn illustrates the potential importance of task assignment procedures. The literature on pay inequality suggests that the properties of the assignment process matter for behavioral responses of workers. While intentional assignments (compared to random assignments) show the clearest behavioral responses (Gächter and Thöni, 2010), justifications of the assignment (e.g. based on performance) might mitigate the effects (Charness and Kuhn, 2007; Breza et al., 2018). Thus, this study provides novel insights on the impact of task assignment procedures on performance and satisfaction.

The rest of this paper is structured as follows. Section 4.2 describes the experimental design, procedures and measurements in more detail. The design and results from the expert survey are depicted in Section 4.3. Section 4.4 presents the data collection procedures and planned analyses.

[2]Participants were separated into different rooms. All people were working on the enriched task and were told that the people in the other room are working on the less enriched task. While this was not the case, it is likely that everybody assumed that there are other people working on the less enriched task. Nonetheless, an ideal experiment would be free of deception (Harrison and List, 2004).

4.2 Research design

4.2.1 Background

Analyzing the impact of task assignment and task comparisons on performance and satisfaction as well as the efficacy of different task assignment procedures relies on a setting with two distinct tasks and the possibility to exogenously and randomly assign a large number of workers to one of these tasks. Another research project I am involved in presented a unique opportunity to investigate this relationship. There, my co-authors and I exogenously vary the existence of an endogenously selected or exogenously assigned leader to analyze the impact of leadership on team performance, team organization and team behavior (Englmaier et al., 2023b).[3] To investigate microaspects of leadership, we collect performance data (completion rates and finishing times), tracking data (using a localization system) and audio data (using voice recorders), as well as detailed data on team member characteristics and creativity (pre-experimental survey). As part of the data analyses, the collected audio data have to be transcribed and evaluated, which are two regular and distinct tasks for research assistants. To transcribe and evaluate these audio data, a large number of research assistants is needed, and to achieve this, students will be invited for a one-time job to assist with the preparation (transcription of audio data) and evaluation (evaluating the transcribed text) of these audio data.[4] This offers the unique opportunity to exogenously vary the task assignment between two close co-workers to analyze the impact of task comparisons and task assignment procedures on performance and satisfaction. According to Harrison and List (2004), this experiment could be categorized as a framed field experiment, since students are explicitly invited to work on the above described tasks (i.e. they do not naturally work on these tasks).

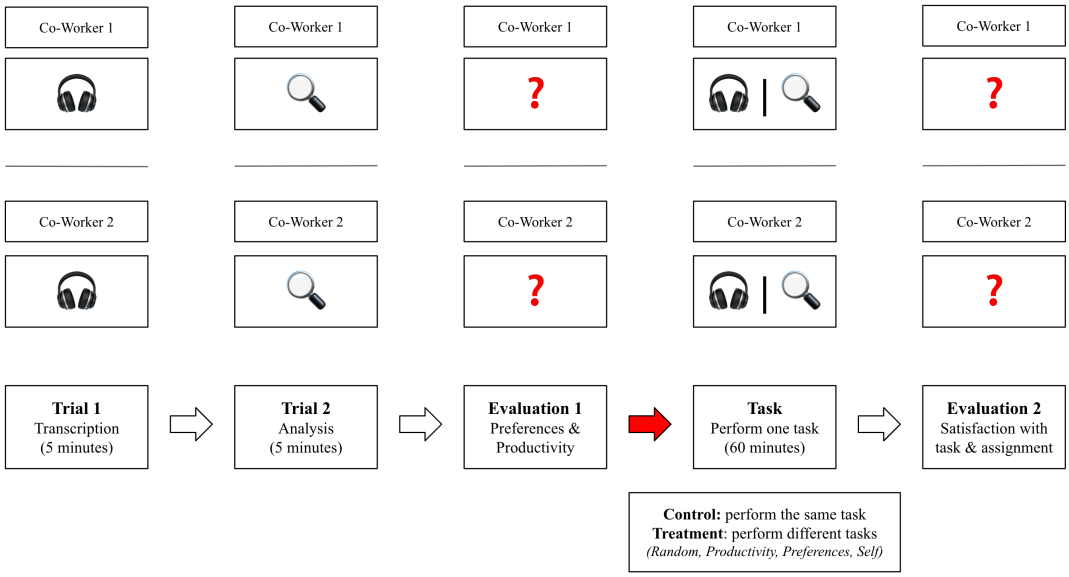
4.2.2 Experimental design

The study will be conducted at the Ludwig-Maximilians-Universität (LMU) in Munich. Participants will be invited in groups of two co-workers via ORSEE (Greiner, 2015) using the subject pool of the MELESSA laboratory at LMU Munich. In the invitation mail,

[3] At the time of writing this chapter, the RCT is still ongoing. Data collection is expected to be completed in April 2023.

[4] In total, we collect audio data for roughly 270 teams with three team members each. A minimum of two research assistants per transcription and evaluation is intended to validate the correctness of the transcriptions and to provide an average evaluation.

Figure 4.1: Experimental design and procedures



Notes: The figure shows different phases of the experiment (co-workers sit across from each other throughout the whole experiment). The headphones represent the transcription task and the magnifying glass represents the analysis task.

participants will be informed that we need assistance for the preparation and analyses of data, that this task will last for roughly 75 minutes and that they will receive a fixed wage of 20 Euro. Furthermore, they are informed that this a one-time job, which excludes potential reputational concerns (see, e.g. Gneezy and List, 2006; Kosfeld and Neckermann, 2011; Kube et al., 2012).

Figure 4.1 presents the experimental design and procedures. Upon arrival, co-workers take a seat across from each other. In a first step, independent of the treatment arm, both workers have to work on both tasks for 5 minutes to get to know them and to illustrate quality requirements. This is followed by a short evaluation of their perceived productivity and preference for both tasks (on a 7-point Likert-scale). Thereafter, both workers are either working on the same task (*Control*) or different tasks (*Treatment*) for 60 minutes.[5] In the latter case, the task assignment procedure is exogenously varied: In treatment *Random*, workers are randomly assigned to tasks. In treatment *Productivity*, workers are assigned according to their perceived productivity, and in treatment *Preferences*, they are assigned according to their preferences for both tasks. In treatment *Self*, workers have the

[5]Participants know who is working on which task. To make the task of the co-worker very prominent, subjects sit across from each other and wear over-ear headphones for the transcription task.

chance to communicate and jointly decide on who is doing which task. After performing the assigned task for 60 minutes, all subjects have to fill out a short survey regarding their satisfaction with the task and the task assignment procedure as well as their level of stress and willingness to perform a similar task again.

4.2.3 Hypotheses

The following hypotheses are derived from a stylized agency model following Englmaier and Leider (2012). There are two different tasks and workers are heterogeneous with respect to productivity (p) and preferences (l) in the respective tasks. The principal hires two workers and pays a fixed wage (w) independent of the provided effort (e). There are no reputational incentives in place since the interaction is one shot. Effort costs are convex and workers that are more productive or have a stronger preference for the task have lower costs. Furthermore, worker's utility depends on how much she cares about (θ) the task assignment, taking the own task (i) and the task of the co-worker (k) into account (with $\theta \in [0, 1]$). The principal's profit is given by $e - w$ and the worker's utility $u(w)$ is given by:

$$u(w) = w + \eta(w - o)(e - w) - \frac{c(e)}{p_i l_i} - \theta(\frac{c(e)}{p_i l_i} - \frac{c(e)}{p_k l_k}) \tag{4.1}$$

where η reflects the worker's reciprocal attitude, which is assumed to be non-negative ($\eta \geq 0$), and o is the outside option. The worker receives a wage gift (w is higher than the outside option o) and thus, her utility increases in the principal's profits. The worker's best response[6] e^* is given by:

$$\frac{\delta u(w)}{\delta e} = \eta(w - o) - \frac{c'(e)}{p_i l_i} - \frac{\theta c'(e)}{p_i l_i} + \frac{\theta c'(e)}{p_k l_k} = 0 \tag{4.2}$$

$$c'(e^*) = \frac{\eta(w - o)p_i l_i p_k l_k}{(1 + \theta)p_k l_k - \theta p_i l_i} \tag{4.3}$$

Assuming that either both workers perform the same task, the worker is equally productive at both tasks and does not prefer one task ($p_i = p_k$ and $l_i = l_k$) or that the worker does not care about the task assignment ($\theta = 0$), the best response e^* is given by:

[6]The first order condition is necessary and sufficient, since the second order condition is globally satisfied for convex costs.

$$c'(e^*) = \eta(w-o)p_i l_i \tag{4.4}$$

Comparing (4.3) and (4.4) immediately leads to the first hypothesis:

Hypothesis 1 *If workers are assigned to different tasks and care about the task assignment, effort is higher (lower) when assigned to the task, which they (do not) prefer or are more (less) productive at.*

Next, I analyze whether the optimal effort choice varies with the productivity and preference for the own (p_i and l_i) and the task of the co-worker (p_k and l_k):

$$\frac{\delta e^*}{\delta p_i} = \frac{\frac{(1+\theta)\eta(w-o)p_k^2 l_k^2 l_i}{[(1+\theta)p_k l_k - \theta p_i l_i]^2}}{c''(e)} > 0 \;; \quad \frac{\delta e^*}{\delta l_i} = \frac{\frac{(1+\theta)\eta(w-o)p_k^2 l_k^2 p_i}{[(1+\theta)p_k l_k - \theta p_i l_i]^2}}{c''(e)} > 0 \tag{4.5}$$

$$\frac{\delta e^*}{\delta p_k} = \frac{\frac{-\theta\eta(w-o)p_i^2 l_i^2 l_k}{[(1+\theta)p_k l_k - \theta p_i l_i]^2}}{c''(e)} < 0 \;; \quad \frac{\delta e^*}{\delta l_k} = \frac{\frac{-\theta\eta(w-o)p_i^2 l_i^2 p_k}{[(1+\theta)p_k l_k - \theta p_i l_i]^2}}{c''(e)} < 0 \tag{4.6}$$

Intuitively, effort increases if a worker is more productive at her own task or prefers her own task more. Accordingly, effort decreases if a worker is more productive at the task of the co-worker or prefers the task of the co-worker more.

Hypothesis 2 *The higher the preference or productivity for the own task (the task of the co-worker), the higher (lower) the provided effort.*

Finally, I analyze whether the optimal effort choice varies with how much a worker cares about the task assignment (θ):

$$\frac{\delta e^*}{\delta \theta} = \frac{\frac{(p_i l_i - p_k l_k)\eta(w-o)p_i l_i p_k l_k}{[(1+\theta)p_k l_k - \theta p_i l_i]^2}}{c''(e)} \tag{4.7}$$

According to (4.7), optimal effort depends on the relationship between $p_i l_i$ and $p_k l_k$, and thus whether the worker works on the task, which she prefers and is more productive at, or on the task, which she does not prefer and is less productive at.

Hypothesis 3 *If workers are assigned to different tasks and work on the task, which they prefer and are more productive at, effort is higher (lower) when they care more (less) about the task assignment.*[7]

Furthermore, I do expect different behavioral responses depending on the task assignment procedure. The literature on wage comparisons suggests that intentional compared to random assignments should lead to the clearest behavioral responses (Gächter and Thöni, 2010). In treatment *Random*, task allocation is exogenously determined and not intentional. Hence, it serves as the cleanest comparison to *Control* and is expected to show the pure impact of task assignment and task comparisons on performance and satisfaction.

Task assignment in treatments *Productivity* and *Preferences* is intentional, but justified by the assignment according to perceived productivity or preferences for both tasks. Thus, both treatments are designed to avoid negative behavioral responses, but their efficacy relies on the acceptance of the respective justification. While productivity is likely to be accepted as a reasonable justification (Charness and Kuhn, 2007; Breza et al., 2018), empirical evidence for the acceptance of (self-evaluated) perceived productivity and preferences as justifications is scarce. Beyond that, even if they are accepted as reasonable justifications, a principal would still limit workers' choice autonomy by dictating the criterion for task assignment. These hidden costs of control could in itself imply negative behavioral responses (Falk and Kosfeld, 2006). Furthermore, Boss et al. (2021) show that autonomy to choose project ideas improves entrepreneurial team performance, which is partly explained by a better match of interests and ideas. Thus, as long as co-workers can agree on a certain task assignment, treatment *Self* is expected to be the most effective task assignment procedure.

In comparison to the other assignment procedures, treatment *Random* is expected to be the least effective, since tasks are only 'efficiently' allocated for every second group of co-workers. However, there are also factors that could lead to inefficient allocations in treatments *Productivity*, *Preferences* and *Self*. While the results in Delfgaauw et al. (2020) suggest strong performance enhancing effects of treatment *Preferences*, causal evidence for the effectiveness of task allocation according to preferences (instead of productivity)

[7] Accordingly, if workers are assigned to different tasks and work on the task, which they do not prefer and are less productive at, effort is lower (higher) when they care more (less) about the task assignment.

is scarce. Treatments *Productivity* and *Self* on the other hand might suffer from inefficient task choices and conflict due to overconfident workers (Köszegi, 2006; Burks et al., 2013), which in turn could lead to inefficient task allocations.[8] However, overconfident behavior of a co-worker is unlikely to further strengthen negative behavioral responses of a worker (Kennedy et al., 2013).

4.2.4 Outcome measures

The focus of the analysis is on the impact of task assignment and task comparisons on performance and satisfaction as well as on the efficacy of different task assignment procedures (i.e. maximizing overall outcomes, while minimizing negative effects). Primary outcome variables (performance and satisfaction) are collected during the task performance stage and the second evaluation phase after participants have performed the task (see Figure 4.1). Performance will be measured quantitatively and qualitatively. For the transcription task, quantity is reflected by the number of words workers have transcribed. Quality will be measured by the number of correctly transcribed words (or inversely, by the number of mistakes). For the evaluation task, quantity is represented by the number of evaluated questions. Quality will be measured by the number of correctly evaluated questions (or inversely, by the number of mistakes).[9] Since performance is measured on different scales, this outcome will be transformed into z-scores by using the *Control* groups' mean and dividing it by the *Control* groups' standard deviation. Accordingly, performance has a mean of zero and a standard deviation of one for both tasks in the *Control* group. Satisfaction will be measured in two dimensions by asking participants how satisfied they were with the task they performed and the task assignment procedure (on a 7-point Likert-scale from 'very dissatisfied' to 'very satisfied').[10]

Besides these primary outcome variables, two short questions on the level of stress and the willingness to perform a similar task again (on a 7-point Likert-scale from 'strongly disagree' to 'strongly agree') are included in the second evaluation phase at the end of the experiment. These secondary outcome measures serve as proxies for hardly observable and potentially long-term consequences including terminations.

[8] Contrasting participants' perceived productivity in the evaluation phase after working on both tasks for 5 minutes with their actual performance serves as a proxy for overconfidence.

[9] Therefore, besides subjective evaluations of team member characteristics and team communication, I also include objectively measurable questions on age, gender, and experience with the task.

[10] The measures for satisfaction will also be transformed into z-scores.

4.3 Expert survey

Leaders face a difficult task when forming expectations about the impact of task assignment and task comparisons on performance and satisfaction. Even if they are aware of the potential influence of task assignment and task comparisons on performance and satisfaction, the question arises which task assignment procedure is the most effective (i.e. maximizes overall outcomes, while minimizing negative effects on workers). To elicit perceptions of actual decision makers and to get a rough estimate of potential effect sizes, I conducted a survey in the run-up to the experiment. In total, I recruited 400 practitioners, who had been involved in personnel decisions as part of their daily work, from the subject pool maintained by the professional survey provider *Cint*.[11]

4.3.1 Survey design

The survey consists of three parts to elicit practitioners' beliefs about the impact of task assignment and task comparisons on performance and satisfaction and the efficacy of different task assignment procedures. Furthermore, this survey seeks to inform about actual task assignment procedures in practice.

Part 1 investigates general perceptions about the impact of task assignment on performance and satisfaction. Participants have to evaluate whether they generally agree with the statement that task assignment has an impact on performance and satisfaction (on a 4-point Likert-scale from 'do not agree' to 'agree').

Part 2 analyzes the expected effect sizes of task assignment and task comparisons on performance and satisfaction and how these effects differ between different task assignment procedures. Participants have to evaluate how large the expected effects (in %) on performance and satisfaction are, when switching from an initial situation in which two co-workers (one female (Anna) and one male (Otto)) work on the same task, to a situation in which these co-workers work on different tasks. Thereby, participants have to differentiate between effects on advantaged (working on the preferred task or the task for which they expect to perform better) and disadvantaged workers (working on the less preferred task or the task for which they expect to perform worse). Participants are

[11]Participants have one of the following primary roles within their organization (profiling option via Cint): Owner or Partner, President/CEO/Chairperson, Middle Management, Chief Financial Officer (CFO), Senior Management, Project Management, Chief Technical Officer (CTO), C-level executive, Director, HR manager. The survey was pre-registered at Aspredicted (#122532).

randomly assigned to one of eight conditions, where I vary the initial situation and who (female/male) is working on which task in the second situation. There are four different initial situations: 1) both workers work on the preferred task, 2) both workers work on the less preferred task, 3) both workers work on the task for which they expect to perform better, 4) both workers work on the task for which they expect to perform worse. In all eight conditions, participants have to evaluate the impact on performance and satisfaction in four cases (different task assignment procedures). In the second situation, tasks are either randomly assigned *(Random)*, assigned according to preferences *(Preferences)*, assigned according to perceived productivity *(Productivity)*, or self-assigned by the workers *(Self)*. The initial situation and the task assignment in the second situation (who is doing which task) stay the same in all four cases.

Part 3 analyzes actual task assignment procedures in practice. Participants should explain how tasks are assigned in their firm (open text) and state whether pre-specified task assignment procedures (ability, preferences, efficiency, self-assignment, seniority, favoritism, fairness) are used in their firm (multiple-choice).[12]

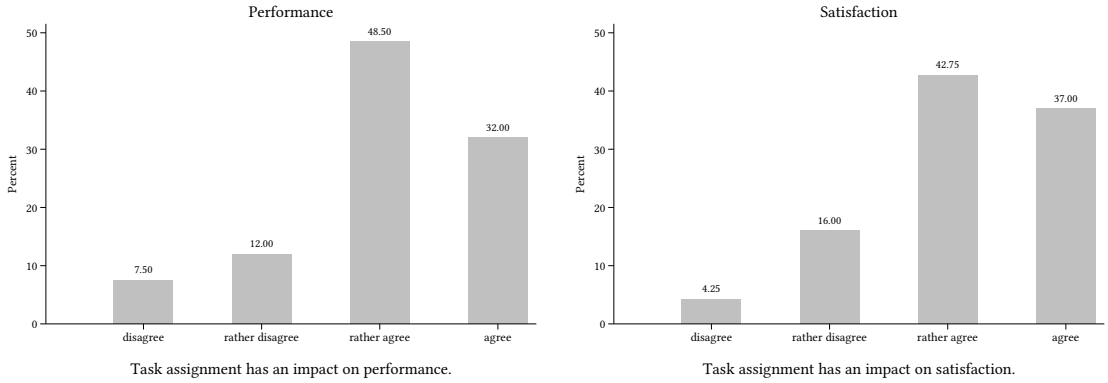
4.3.2 Survey results

First and foremost, practitioners clearly expect an impact of task assignment on performance and satisfaction (see Figure 4.2). Roughly 80 percent of practitioners rather agree or agree, while less than 10 percent disagree with the statements that task assignment has an impact on performance and satisfaction of employees. Thus, in general, they consider task assignment to be an important factor for performance and satisfaction.

These findings are further confirmed by expected effect sizes (Part 2) on performance (see Figure 4.3) and satisfaction (see Figure 4.4). The distribution of expected effect sizes, when moving away from a situation in which both co-workers work on the same task, to a situation in which they work on different tasks, shows three interesting patterns. First, it is noticeable that only a very small fraction of practitioners expect no effect of task comparisons on performance and satisfaction, independent of whether the worker is advantaged (working on the preferred task or the task for which she expects to perform better) or disadvantaged (working on the less preferred task or the task for which she expects to perform worse).

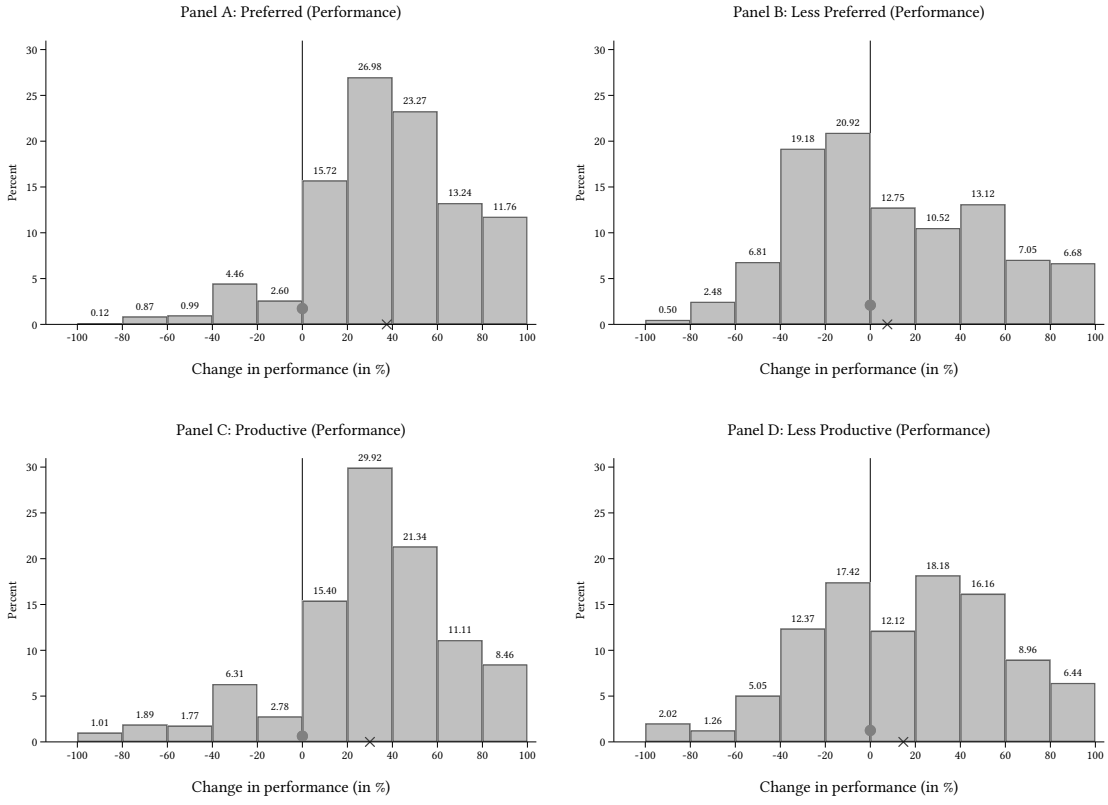
[12]The selection is inspired by Delfgaauw et al. (2020), while I have further added self-assignment as a potential procedure of task assignment.

Figure 4.2: Impact of task assignment on performance and satisfaction



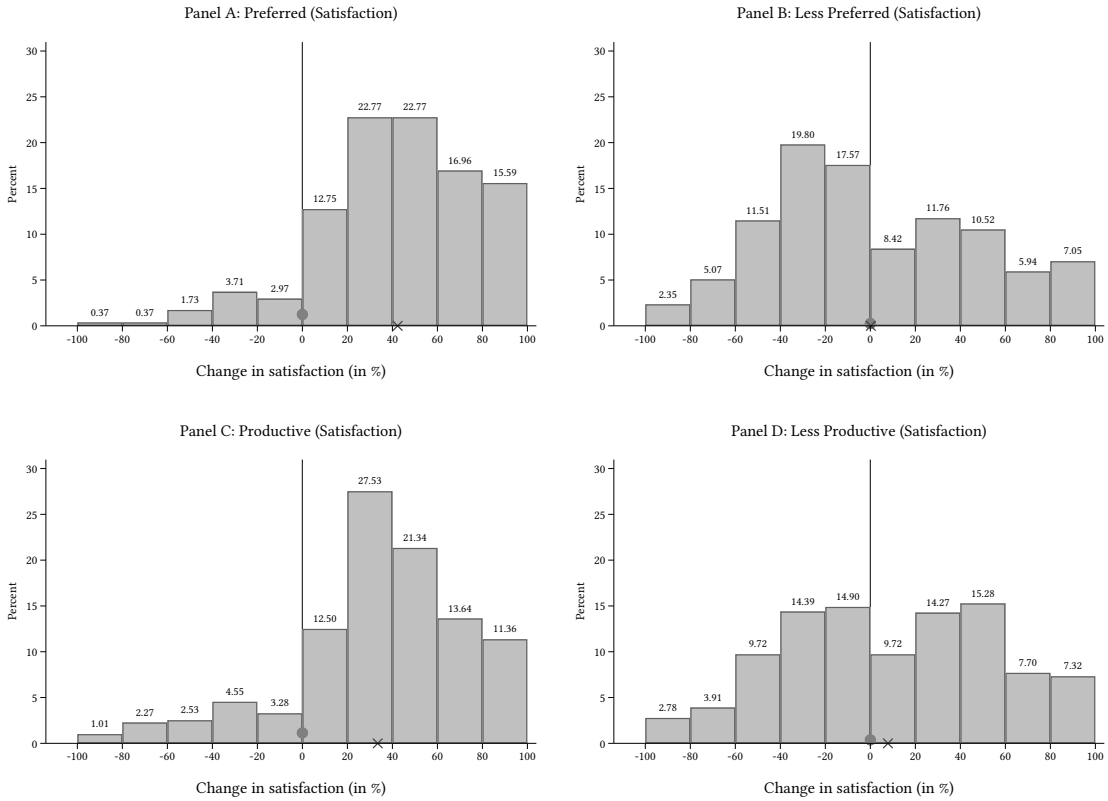
Notes: The figure shows histograms of survey answers on whether participants believe that task assignment has an impact on performance and satisfaction of employees. For each of the two statements, subjects had to evaluate how much they agree on a 4-point Likert-scale from 'disagree' to 'agree'.

Figure 4.3: Expected effect sizes on performance



Notes: The figure shows histograms of expected effect sizes (in %) of task assignment and task comparisons on performance. Panel A (Panel C) shows the distribution of expected effect sizes for workers that work on the preferred task (the task for which they expect to perform better). Panel B (Panel D) shows the distribution of expected effect sizes for workers that work on the less preferred task (the task for which they expect to perform worse). The circle represents the share of practitioners that expect no effect on performance and the 'X' represents the mean of the expected effect size.

Figure 4.4: Expected effect sizes on satisfaction



Notes: The figure shows histograms of expected effect sizes (in %) of task assignment and task comparisons on satisfaction. Panel A (Panel C) shows the distribution of expected effect sizes for workers that work on the preferred task (the task for which they expect to perform better). Panel B (Panel D) shows the distribution of expected effect sizes for workers that work on the less preferred task (the task for which they expect to perform worse). The circle represents the share of practitioners that expect no effect on satisfaction and the 'X' represents the mean of the expected effect size.

Second, they expect on average large positive effects on performance and satisfaction for advantaged workers (working on the preferred task or the task for which they expect to perform better). This is also confirmed by the results from Probit regressions of the probability of a decrease or an increase, and OLS regressions of the overall change of performance and satisfaction (see Appendix Table 4.A.1). The probability of a decrease (an increase) in performance and satisfaction is expected to be much lower (higher) when working on the preferred task or the task for which they expect to perform better. Accordingly, the overall expected change of performance and satisfaction is significantly larger for advantaged workers. Comparing the expected effects between the preferred task and the task for which a worker expects to perform better further illustrates that overall effects are expected to be larger when focussing on preferences (lower probability of a decrease, higher probability of an increase, larger positive change).

Third, expectations regarding the impact of task assignment on performance and satisfaction of disadvantaged workers (working on the less preferred task or the task for which they expect to perform worse) are mixed. When switching from working on the same task to working on the less preferred task (while the co-worker is working on the preferred task), practitioners expect a performance decrease in roughly half of the situations (49.9%), while they expect a performance increase for almost all other situations (48.0%). The impact on satisfaction tends to be even more negative (56.3% decrease, 43.4% increase). When focussing on perceived productivity, the effects are expected to be less pronounced. Less practitioners expect a decrease, more practitioners expect an increase, and accordingly, they also expect a larger positive overall change.

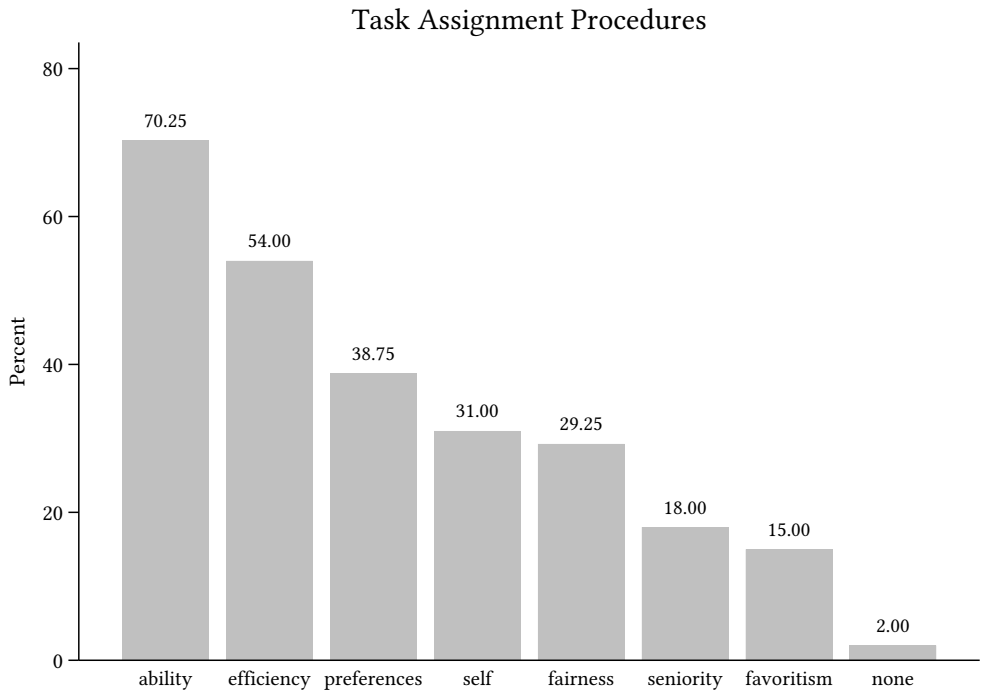
While negative effects on performance and satisfaction are in line with the predictions of the model in Section 4.2, positive expectations could be driven by lower comparability between co-workers when working on different tasks. This in turn could lead to less pressure (Bellemare et al., 2010) and competition among team members (Gneezy et al., 2003; Gneezy and Rustichini, 2004; Niederle and Vesterlund, 2007).[13] In the experiment, subjects are informed that they are working on different audio and transcribed files, making them aware that perfect comparability is impossible. Another potential explanation for positive effects is that workers might feel less responsible and shirk when working on the same task as their co-workers. In this context, Maximiano et al. (2007) show that effort levels are only marginally lower in multi-worker compared to one-worker firms. Nonetheless, the importance of every single transcript and evaluation is emphasized to increase the perceived responsibility of workers in all situations.

Furthermore, practitioners do not expect vastly different effect sizes for the different task assignment procedures (see Appendix Table 4.A.2 and Table 4.A.3). If anything, they expect treatment *Self* to be accompanied by higher levels of performance and satisfaction for workers that work on the less preferred task. However, given similar expected effect sizes, the question arises whether the different task assignment procedures are similarly applied within firms.

The last part of the survey seeks to inform about actual task assignment procedures in practice. Figure 4.5 shows how many practitioners stated that the respective task assignment procedure is used in their firm. The by far most applied task assignment proce-

[13] This would also be in line with the positive coefficient of female workers on performance, particularly when there is a focus on perceived productivity (see Appendix Table 4.A.2).

Figure 4.5: Task assignment procedures in practice



Notes: The figure shows the share of participants that stated that the respective task assignment procedure is used in their firm.

dure is allocating according to ability (70.25%). Roughly half of the participants (54.00%) state that tasks are assigned for reasons of efficiency. 38.75% state that they assign tasks according to preferences, 31.00% use self-assignment as an assignment procedure and 29.25% allocate tasks with respect to fairness concerns. A small share states that tasks are assigned according to seniority (18.00%) or out of favoritism (15.00%). Almost nobody (2.00%) stated that none of the other assignment procedures is used in their firm.[14]

Taken together, the survey highlights the heterogeneous beliefs of practitioners about the impact of task assignment and task comparisons on performance and satisfaction, especially with respect to the impact on disadvantaged workers. Furthermore, while they expect no large differences between the different task assignment procedures, there is a considerable difference in the application of these procedures in practice. Thus, practitioners might either misjudge the efficacy of the different task assignment procedures or make inefficient use of some of these, since especially self-assignment could poten-

[14]These results are also in line with statements in the open text form (see Appendix Figure 4.A.1). The by far most used words relate to ability, knowledge and competence.

tially reduce time-consuming considerations of managers. Beyond that, roughly half of the participants believe that tasks are not assigned for reasons of efficiency, which leaves room for improvement and underpins the importance of empirical evidence analyzing the impact of task assignment, task comparisons and different task assignment procedures on performance and satisfaction.

4.4 Data and analysis

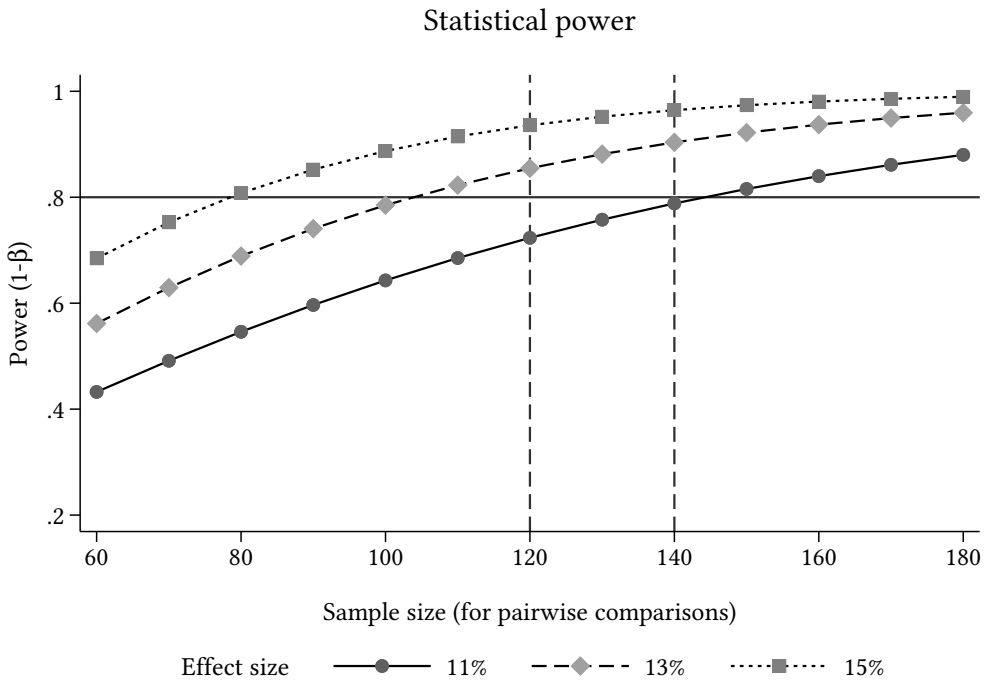
4.4.1 Data collection

This experiment uses audio records from teams participating in an ongoing RCT, which eventually contain information about the performed task in this RCT (Englmaier et al., 2023b). Thus, to avoid contamination, since subjects potentially participate in both studies, the experiment will only start after data collection has been completed. Thereafter, subjects will be invited to this study in groups of two co-workers, which will work in one of two different rooms. I plan to conduct five sessions (per room) per day and thus collect data for ten groups of two co-workers per day (twenty individuals in total and ten individuals per task). I implement five experimental variations, which I randomize on a daily level to avoid treatment spillovers. The script for the implementation of treatments and the instructions for both tasks are depicted in Appendix Section 4.A.2.

4.4.2 Statistical power and sample size

The expectations regarding the effects of task comparisons on performance are heterogeneous. While the results in Montagno (1985) already suggest large positive effects (11% increase in quantity and 22% decrease in errors), practitioners expect on average even larger effects (more than 30% increase) on the performance of advantaged workers (working on the preferred task or the task for which they expect to perform better). I also expect effects on advantaged workers to be larger than in Montagno (1985), since co-workers sit across from each other and can observe which task a close co-worker is working on, instead of referring to some other worker. Nonetheless, power is calculated for different effect sizes, including an effect size of 11%, and different sample sizes (for pairwise comparisons between different conditions). Power calculations (see Figure 4.6) indicate that with a sample size of 120 to 140 subjects (for pairwise comparisons), the de-

Figure 4.6: Statistical power and sample size



Notes: The figure shows the statistical power to detect an effect of 11%, 13%, or 15% for different sample sizes (for pairwise comparisons). The sample size illustrates the sum of individuals working on one of the two tasks in two different conditions.

sign would be sufficiently powerful (72% to 79%) to detect an effect size of 11%.[15] Due to the wide range of expected effects on performance and to get an order of magnitude for the expected effects on the other outcome measures, I will conduct a pilot with twenty subjects per condition (ten subjects per task per condition).[16]

4.4.3 Analysis

First, I will establish the causal effect of task assignment and task comparisons on the primary outcome variables performance and satisfaction. In treatment *Random*, task allocation is exogenously determined and thus it serves as the cleanest comparison to *Control* to determine the effects of interest. For all other assignment procedures, it is endogenously

[15]Power is calculated based on the performance of five research assistants in the transcription task. I expect the effects on disadvantaged workers (working on the less preferred task or the task for which they expect to perform worse) to be negative, but similar or even larger in magnitude than for advantaged workers. For an effect size of 11%, a sample size of 120 to 140 subjects per condition (60 to 70 subjects per task per condition) would lead to a sufficiently powerful design, ending up with 600 to 700 subjects in total.

[16]The pilot is pre-registered at Aspredicted (#125298).

determined who is working on which task. Therefore, I will focus on comparisons between *Control* and treatment *Random* for the first part of the analysis. As described above, performance is measured quantitatively and qualitatively, and satisfaction is measured on a 7-point Likert-scale. Both outcome measures will be standardized using the *Control* groups' mean and dividing it by the *Control* groups' standard deviation.[17] I will further analyze the impact of task assignment and task comparisons on the secondary outcome variables perceived stress and willingness to perform a similar task again. I will separately estimate the effect on the outcome variables using OLS regressions according to the following equation:[18]

$$Y_i = \beta_0 + \beta_1 \, \textit{Pref}_i + \beta_2 \, \textit{Diff}_i + \beta_3 \, \textit{Pref}_i \times \textit{Diff}_i + \epsilon_i \tag{4.8}$$

where Y_i is the respective outcome for individual i. The indicator variable $\textit{Pref}_i$ ($\textit{Prod}_i$) is equal to one if individual i was randomly assigned to work on the preferred task (the task for which she expects to perform better) and zero otherwise. $\textit{Diff}_i$ is equal to one if the co-worker is working on the other task (treatment *Random*) and zero otherwise. $\textit{Pref}_i \times \textit{Diff}_i$ ($\textit{Prod}_i \times \textit{Diff}_i$) represents the interaction term: Individual i works on the preferred task (the task for which she expects to perform better), while her co-worker is working on the less preferred task (the task for which she expects to perform worse). I will also assess whether the effects differ by gender.

Second, I will analyze the impact of perceived productivity and preferences for the own task and the task of the co-worker on this relationship. I will separately estimate the effect on the outcome variables using OLS regressions according to the following equations:[19]

$$\begin{aligned} Y_i = \beta_0 &+ \beta_1 \, \textit{Pref}_i + \beta_2 \, \textit{Diff}_i + \beta_3 \, \textit{Pref}_i \times \textit{Diff}_i + \beta_4 \, \textit{PrefOwn}_i \\ &+ \beta_5 \, \textit{PrefOwn}_i \times \textit{Pref}_i + \beta_6 \, \textit{PrefOwn}_i \times \textit{Diff}_i \\ &+ \beta_7 \, \textit{PrefOwn}_i \times \textit{Pref}_i \times \textit{Diff}_i + \epsilon_i \end{aligned} \tag{4.9}$$

[17] Performance is standardized for each task separately.

[18] I will run the same regressions with $\textit{Prod}_i$ instead of $\textit{Pref}_i$.

[19] I will run the same regressions with $\textit{ProdOwn}_i$ and $\textit{ProdOther}_i$ instead of $\textit{PrefOwn}_i$ and $\textit{PrefOther}_i$. Furthermore, I will also separately estimate the effect for workers that work on the preferred task or the task for which they expect to perform better and workers that work on the less preferred task or the task for which they expect to perform worse.

$$\begin{aligned} Y_i = \beta_0 &+ \beta_1 \, \textit{Pref}_i + \beta_2 \, \textit{Diff}_i + \beta_3 \, \textit{Pref}_i \times \textit{Diff}_i + \beta_4 \, \textit{PrefOther}_i \\ &+ \beta_5 \, \textit{PrefOther}_i \times \textit{Pref}_i + \beta_6 \, \textit{PrefOther}_i \times \textit{Diff}_i \\ &+ \beta_7 \, \textit{PrefOther}_i \times \textit{Pref}_i \times \textit{Diff}_i + \epsilon_i \end{aligned} \tag{4.10}$$

where Y_i is again the respective outcome for individual i. $\textit{PrefOwn}_i$ and $\textit{PrefOther}_i$ ($\textit{ProdOwn}_i$ and $\textit{ProdOther}_i$) reflect the preferences (perceived productivity) of individual i for her own task and the task of her co-worker (measured on a 7-point Likert-scale).

Third, I will analyze the efficacy of different task assignment procedures for team performance and satisfaction as well as for the secondary outcome variables (for each team of two co-workers). I will separately estimate the effect on the outcome variables using OLS regressions according to the following equation:

$$Y_t = \beta_0 + \beta_1 \, \textit{Treat}_t + \epsilon_t \tag{4.11}$$

where Y_t is the respective outcome for team t. $\textit{Treat}_t$ is a treatment indicator and reflects the different task assignment procedures (with treatment *Random* as the base category). I will also assess whether the effects differ by gender composition. I will further analyze the impact of different task assignment procedures on individual performance and satisfaction and secondary outcome variables. I will separately estimate the effect on the outcome variables using OLS regressions according to the following equation:[20]

$$Y_i = \beta_0 + \beta_1 \, \textit{Pref}_i + \beta_2 \, \textit{Treat}_i + \beta_3 \, \textit{Pref}_i \times \textit{Treat}_i + \epsilon_i \tag{4.12}$$

where Y_i is again the respective outcome for individual i. The indicator variable $\textit{Pref}_i$ ($\textit{Prod}_i$) is equal to one if individual i works on the preferred task (the task for which she expects to perform better) and zero otherwise. $\textit{Treat}_i$ is a treatment indicator and reflects the task assignment procedure (with *Random* as the base category). $\textit{Pref}_i \times \textit{Treat}_i$ ($\textit{Prod}_i \times \textit{Treat}_i$) represents the interaction term: Individual i works on the preferred task (the task for which she expects to perform better) in $\textit{Treat}_i$ (compared to *Random*), while her co-worker is working on the less preferred task (the task for which she expects to perform worse). I will also assess whether the effects differ by gender.

[20] I will run the same regressions with $\textit{Prod}_i$ instead of $\textit{Pref}_i$.

Appendices

Appendix to Chapter 1

1.A.1 Screenshot of an actual ranking on Facebook

Figure 1.A.1: Screenshot of an actual ranking on Facebook (in German)

Notes: The figure shows a screenshot of an actual ranking on Facebook (in German). Teams are ranked according to their finishing times and all teams that did not complete the task are assigned to the same rank.

1.A.2 Direct treatment comparisons

Table 1.A.1 shows summary statistics of the probability of completion, finishing time, number of hints and the probability of purchasing a voucher. Complementing our main analyses, which compares each subsequent component to treatments including the prior ones, Table 1.A.2 compares each treatment directly to *Control*. By design, the results for treatment *T1 (Identity)* remain the same. Comparing *T2 (Identity, Rank)* to *Control*, we see that *T2 (Identity, Rank)* increases completion rates and lowers finishing times on average, but not significantly so, due to heterogeneity in reactions to the ranking (see Panel B in Figure 1.1). As compared to *Control*, Treatment *T3 (Identity, Rank, Prize)* significantly increases the likelihood of succeeding within 60 minutes and significantly lowers finishing times across all four specifications. The completion rate increases by more than 20 percent (almost 12 percentage points) and the remaining time is almost doubled (more than 3 minutes lower finishing times). Further, we provide alternative specifications using linear regressions and GLM models with log link in Table 1.A.3, confirming the robustness of these findings. Since the salience of team identity is an innate feature of tournaments, Tables 1.A.2 and 1.A.3 further provide p-values from Wald tests for the differences between *T1 (Identity)* and *T2 (Identity, Rank)*, and *T1 (Identity)* and *T3 (Identity, Rank, Prize)* (see rows 4 to 6). Akin to business contexts in which team identity is already salient (e.g., due to existing names for the team or brand), this comparison reveals the effects of *T2 (Identity, Rank* and *T3 (Identity, Rank, Prize)* when teams have an identity-related team name. These comparisons reveal that, on average, *T2 (Identity, Rank)* significantly improves teams' finishing times (see specifications (5) to (8)) as compared to *T1 (Identity)* ($0.028 < p < 0.078$), and *T3 (Identity, Rank, Prize)* improves both the likelihood of completion (specifications (1) to (4)) as well as finishing times (specifications (5) to (8), $0.004 < p < 0.039$).

Table 1.A.1: Summary statistics

	Control - Mean (SD)	*T1* *Identity* Mean (SD)	*T2* *Identity, Rank* Mean (SD)	*T3* *Identity, Rank, Prize* Mean (SD)
Completion	0.53 (0.50)	0.42 (0.50)	0.56 (0.50)	0.65 (0.48)
Finishing time	56.47 (5.49)	56.93 (5.32)	55.03 (6.44)	53.80 (7.57)
Number of hints	3.39 (1.35)	3.36 (1.40)	3.31 (1.31)	3.18 (1.29)
Purchased a voucher	0.18 (0.38)	0.20 (0.43)	0.28 (0.52)	0.30 (0.66)
Observations	112	85	94	82

Notes: Completion denotes the share of teams that managed to complete the task within the given time limit of 60 minutes. Finishing time denotes the average time to complete the task (all teams that did not manage to complete the task within 60 minutes are assigned a finishing time of 60 minutes). Number of hints denotes the average number of hints teams' took. Purchased a voucher denotes the share of teams that purchased a voucher for future participation (at a reduced rate). Suggestive differences in completion probability between T1 and *Control* and voucher purchases in T3 and T2 vs. T1 and *Control* are not statistically significant in any of the regression results in which we correct for the influence of potential confounders in the form of fixed effects and control variables.

Table 1.A.2: Team performance (completion and finishing times)

	Completed within 60 minutes				Finishing time			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
T1	-0.086	-0.099	-0.048	-0.045	1.377	1.910	1.668	1.590
Identity	(0.052)	(0.066)	(0.060)	(0.056)	(0.870)	(0.970)	(1.180)	(1.117)
	[0.198]	[0.241]	[0.434]	[0.447]	[0.219]	[0.145]	[0.206]	[0.218]
T2	0.019	-0.005	0.033	0.034	-1.411	-0.673	-0.906	-0.925
Identity, Rank	(0.045)	(0.050)	(0.050)	(0.041)	(0.883)	(0.892)	(1.082)	(1.008)
	[0.690]	[0.918]	[0.524]	[0.481]	[0.214]	[0.484]	[0.437]	[0.396]
T3	0.110**	0.087*	0.112**	0.118**	-3.625**	-3.064**	-3.106**	-3.255**
Identity, Rank, Prize	(0.038)	(0.045)	(0.044)	(0.043)	(1.026)	(1.320)	(1.333)	(1.349)
	[0.019]	[0.080]	[0.032]	[0.026]	[0.033]	[0.039]	[0.036]	[0.032]
T1 = T2	[0.126]	[0.182]	[0.230]	[0.188]	[0.055]	[0.051]	[0.034]	[0.034]
T2 = T3	[0.047]	[0.032]	[0.033]	[0.020]	[0.042]	[0.033]	[0.040]	[0.064]
T1 = T3	[0.026]	[0.039]	[0.025]	[0.030]	[0.012]	[0.011]	[0.008]	[0.013]
Mean in Control	0.527	0.527	0.527	0.527	56.470	56.470	56.470	56.470
Observations	373	373	373	373	373	373	373	373
Team Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Staff FE	No	No	Yes	Yes	No	No	Yes	Yes
Weekday 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)). All columns include room fixed effects. Each column indicates whether team controls (group size, share of males, experience, median age, language, private), staff, and weekday fixed effects are included. Standard errors in parentheses are clustered at the week level. *p*-values from score bootstrapping following Kline and Santos (2012) are listed in square brackets, with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

Table 1.A.3: Team performance (completion and finishing times)

	Completed within 60 minutes				Finishing time			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
T1	-0.091	-0.102	-0.050	-0.048	0.006	0.013	0.015	0.016
Identity	(0.052)	(0.066)	(0.061)	(0.057)	(0.009)	(0.010)	(0.012)	(0.012)
	[0.210]	[0.261]	[0.522]	[0.504]	[0.538]	[0.231]	[0.254]	[0.272]
T2	0.014	-0.006	0.033	0.035	-0.019	-0.012	-0.011	-0.012
Identity, Rank	(0.043)	(0.049)	(0.049)	(0.039)	(0.011)	(0.010)	(0.012)	(0.013)
	[0.819]	[0.908]	[0.586]	[0.452]	[0.165]	[0.278]	[0.407]	[0.389]
T3	0.105**	0.092**	0.118**	0.121***	-0.041**	-0.034*	-0.032*	-0.034*
Identity, Rank, Prize	(0.037)	(0.043)	(0.043)	(0.042)	(0.014)	(0.017)	(0.016)	(0.017)
	[0.039]	[0.043]	[0.013]	[0.010]	[0.033]	[0.089]	[0.058]	[0.087]
T1 = T2	[0.126]	[0.177]	[0.291]	[0.259]	[0.078]	[0.046]	[0.032]	[0.028]
T2 = T3	[0.044]	[0.026]	[0.078]	[0.030]	[0.146]	[0.220]	[0.266]	[0.248]
T1 = T3	[0.036]	[0.037]	[0.037]	[0.027]	[0.004]	[0.012]	[0.017]	[0.031]
Mean in Control	0.527	0.527	0.527	0.527	56.470	56.470	56.470	56.470
Observations	373	373	373	373	373	373	373	373
Team Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Staff FE	No	No	Yes	Yes	No	No	Yes	Yes
Weekday FE	No	No	No	Yes	No	No	No	Yes

Notes: The table displays average marginal effects 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)). All columns include room fixed effects. Each column indicates whether team controls (group size, share of males, experience, median age, language, private), staff, and weekday fixed effects are included. Standard errors in parentheses are clustered at the week level. *p*-values from score bootstrapping following Kline and Santos (2012) are listed in square brackets, with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

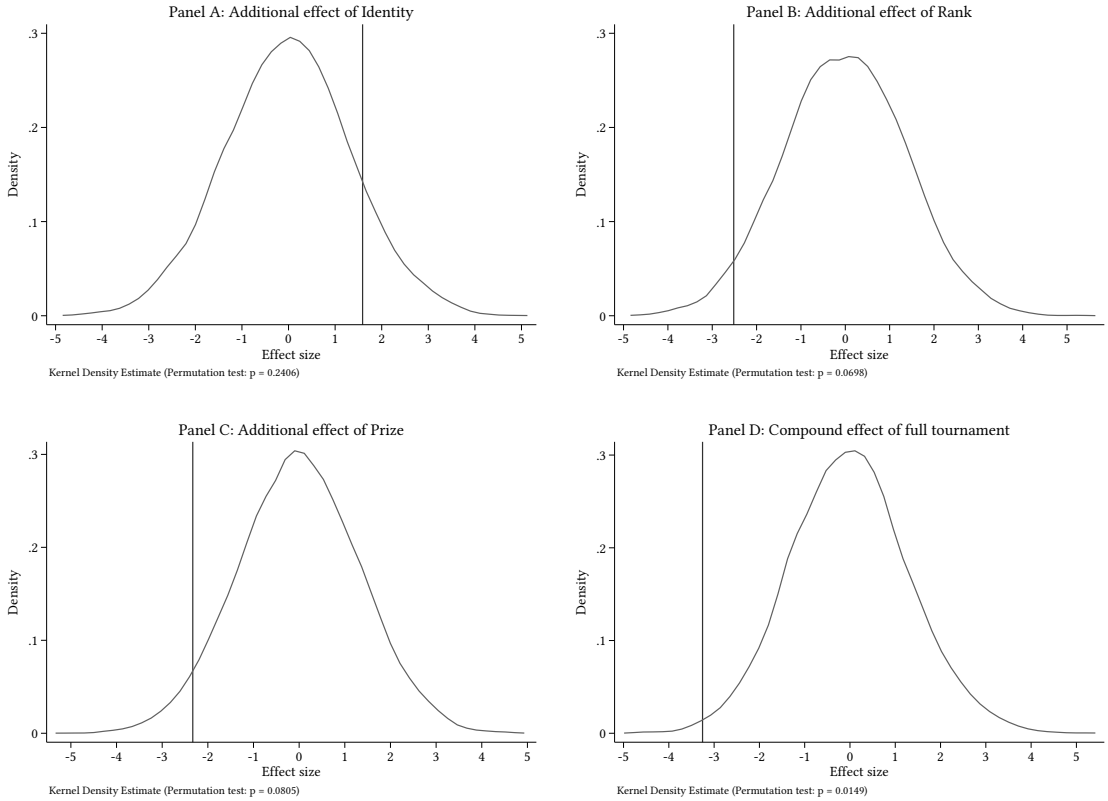
1.A.3 Randomization inference

In addition, we have also carried out a randomization inference exercise (Athey and Imbens, 2017). Because a treatment effect may also arise due to the randomness of who gets assigned to which condition, we want to establish the probability that our findings indeed result from the treatment. Intuitively, randomization inference asks what would have occurred not only under the actual random assignment, but whether the result would also hold under all possible random assignments of treatments to data. We randomly assigned treatment status (preserving the original ratio between treatments) to observations and estimated our regression equation of interest. By repeating this procedure 10,000 times, we obtain a distribution of counterfactual estimates to which we can compare our actual estimates. The resulting randomization inference p-value is equivalent to the proportion of times the placebo treatment effect was more extreme than the estimated actual treatment effect.

As in our main analyses in Section 1.3.1, we focus on comparing each "subsequent" treatment group to the "prior" one using dummy variables for each added component. Figure 1.A.2 plots the randomization distributions of the effect sizes of adding *Identity* (Panel A), adding *Rank* (Panel B), adding *Prize* (Panel C) and the overall effect of *T3 (Identity, Rank, Prize)* relative to *Control* (Panel D) on finishing time. We abstain from a randomization inference exercise on the probability of finishing the task, because the necessary additivity assumption for constructing a confidence interval is unlikely to be fulfilled for binary outcome variables (Rigdon and Hudgens, 2015).

In each panel, the vertical, solid lines indicate the actually observed effect. Panel A shows that the true effect of *Identity* does not appear extreme, and with $p = 0.2406$, we cannot reject the null hypothesis of no individual effect. This is different in Panel B, where we plot the distributions for teams that are subjected to a ranking (*Rank*) in addition. With $p = 0.0698$, the true effect of a reduced finishing time seems unlikely to be a statistical artefact. Panel C shows the randomization distribution for teams with the additional opportunity to win a monetary prize (on top of being ranked). These teams are much quicker than a random distribution of treatments across observations would have suggested ($p = 0.0805$). Lastly, Panel D shows the randomization distribution for the overall effect of *T3 (Identity, Rank, Prize)* compared to *Control*. The result supports our finding that a tournament with a monetary prize and a ranking of teams by their team

Figure 1.A.2: Randomization distributions of effect sizes



Notes: The figure plots the randomization distributions (10,000 resampling replications) of finishing times. The vertical line in each graph shows the observed effect size for adding *Identity* (Panel A), adding *Rank* (Panel B), adding *Prize* (Panel C), or for adding all tournament features simultaneously (Panel D).

name reduces finishing times substantially ($p = 0.0149$). To summarize, all four panels show that our previous results are robust to randomization inference.

1.A.4 Further heterogeneity analyses

Providing an alternative specification, Figure 1.A.3 shows quantile regressions on residualized finishing times using a fully specified GLM regression (with log link). The results are similar (compared to Figure 1.1), confirming the robustness of these findings.

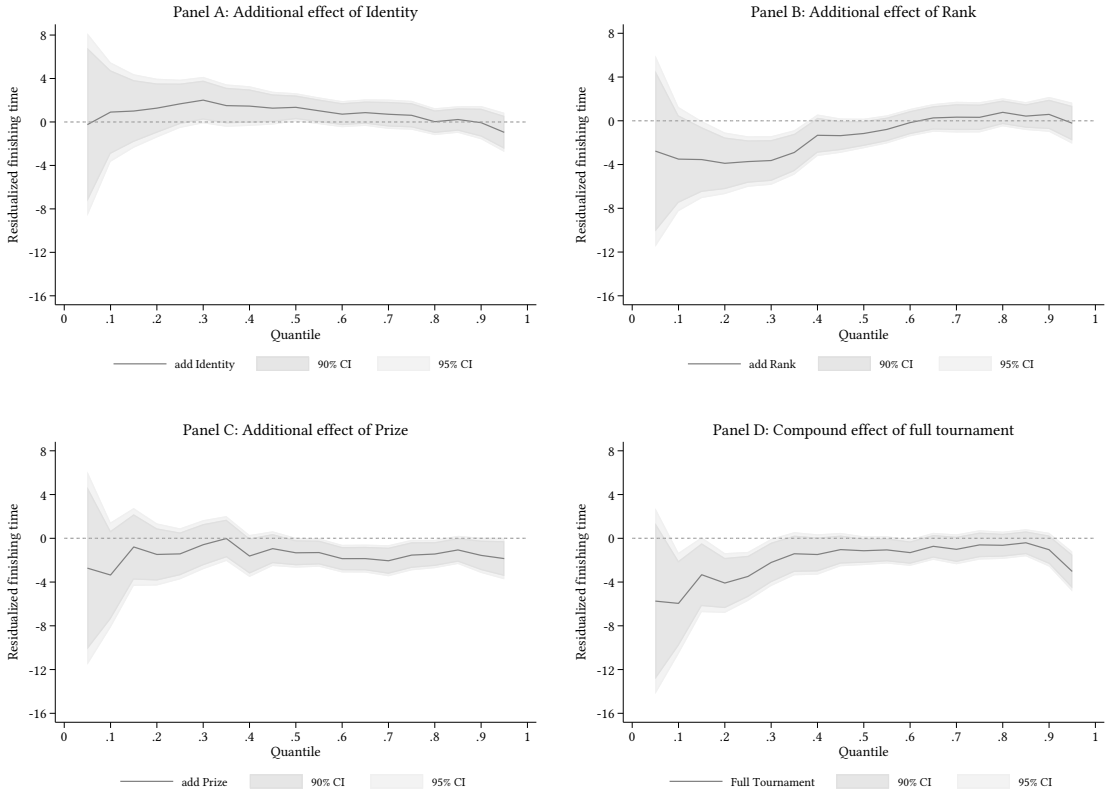
To understand the importance of the composition of a team for possible heterogeneity in the observed treatment effects, we estimate whether (and how fast) teams finish the task in linear probability (and Tobit) models by including interaction terms between each treatment component (i.e., the dummy variables *Identity*, *Rank*, and *Prize*) and observable team characteristics. Appendix Table 1.A.4 shows that, for the probability of completing the task, adding a ranking (*Rank*) interacts positively with the share of males in a team. This is not only in line with the recent literature on gender differences in the willingness to compete (Niederle and Vesterlund, 2011), but also with recent evidence from laboratory experiments studying the role of gender in individual competition without prizes in a routine task (Schram et al., 2019). In contrast, introducing a *Prize* in addition to the ranking tends to increase team performance irrespective of the observed gender composition and other team characteristics. The latter provides suggestive evidence for agency theory (irrespective of gender) from individual (and mostly routine) tasks (see Bandiera et al., 2021).

Tobit regressions on finishing times as reported in Table 1.A.5 yield results in line with the above-mentioned interaction effect for *Rank*, although less precisely estimated. The more males there are in a team, the stronger the reduction in finishing times due to the competition introduced in *Rank*. Further, they reveal a more nuanced picture in terms of image and instrumental concerns. It turns out that the image concerns prevalent in *Rank* are particularly effective in reducing the finishing times of teams that performed the task with their colleagues (company booking), whereas the additional monetary incentive in *Prize* was particularly effective in stimulating the performance of private teams (regular booking).

One reason for the differential treatment effect of prizes for groups of colleagues could be driven by pessimistic expectations about the sharing norm among company team members, who might expect not to be able to receive a fair share of the prize. To explore this argument, we conducted an additional survey in which we elicited social norms of prize sharing following the incentivized elicitation procedure of Krupka and

Weber (2013). We recruited an online sample ($n = 209$) of subjects that had experience with real life escape challenges. We asked them about the appropriateness of different sharing norms across five scenarios. All scenarios were based on the situation in the actual escape challenge (in which winning teams of a given week were informed that they could send a team member to collect their prize money) and we varied how the prize was shared across the five scenarios. For each scenario and in randomized order, subjects had to evaluate the social appropriateness of how the prize is shared within a group of friends (taking part in their leisure time) or a group of colleagues (taking part in a team-building event) on a 4-point Likert-scale from "very socially inappropriate" to "very socially appropriate". One in one hundred participants was eligible for an additional payment and participants were informed that if they choose the same answer as the majority of all other survey participants in one randomly selected scenario, they would earn 50 Euro (if they were randomly selected for payment). The histograms in Figure 1.A.4 show that the equal sharing norm is considered most appropriate, and, more importantly, that there are no systematic differences in sharing norms across types of teams. χ^2-tests comparing responses regarding groups of friends vs. groups of colleagues within each scenario cannot reject the equality of underlying distributions (p-values in brackets). *Scenario 1:* "The person who collects the prize receives all of the prize money (150 Euro)." (p-value: 0.986). *Scenario 2:* "The prize money (150 Euro) will be divided equally among all members of the group." (p-value: 0.699). *Scenario 3:* "The prize money (150 Euro) will be divided unequally among all members of the group." (p-value: 0.681). *Scenario 4:* "The person who collects the prize receives half of the prize money (75 Euro) and the rest will be divided equally among all members of the group." (p-value: 0.937). *Scenario 5:* "The person who collects the prize receives half of the prize money (75 Euro) and the rest will be divided unequally among all members of the group." (p-value: 0.783). Hence, differences in expected sharing norms are unlikely to explain differential treatment effects across private and company teams. Of course, there exist several other potential explanations for differences in the observed coefficients for private and company teams. First, a primary reason for company teams to face the escape challenge may be bonding purposes as part of a team building event, which may render additional monetary incentives less effective. Second, there could be differences in income or wealth between teams of friends and teams of colleagues, that affect the perceived size of the incentive. Third, company teams may have formed less optimistic expectations about their subjective probability of winning

Figure 1.A.3: Quantile regressions on residualized finishing times



Notes: The figure shows quantile regressions on residualized finishing times. Panel A shows the additional effect of salient team identity. Panel B shows the additional effect of a public ranking. Panel C shows the additional effect of a monetary prize. And Panel D shows the overall effect of a tournament with a monetary prize (compares *T3 (Identity, Rank, Prize)* to *Control*). The line at zero marks residualized finishing times in the comparison group. Negative (positive) values indicate reductions (increases) in residualized finishing times due to *Identity* (Panel A), *Rank* (Panel B), *Prize* (Panel C), or due to adding all tournament features simultaneously (Panel D).

the prize and therefore reacted less to incentives. As we observe only a relatively small number of team-building event groups in our sample, we see scope for future research on this exploratory finding and the potential additional channels discussed above.

Table 1.A.4: Team performance (completion, interactions)

	Completed within 60 minutes						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
+ Identity	-0.048	0.167	-0.133	-0.211	-0.071	0.008	-0.715
(making identity salient)	(0.057)	(0.344)	(0.143)	(0.182)	(0.130)	(0.105)	(0.151)
	[0.504]	[0.651]	[0.444]	[0.588]	[0.676]	[0.946]	[0.294]
+ Rank	0.083	-0.264	0.169	0.224	-0.066	-0.109	0.409
(adding a ranking)	(0.047)	(0.294)	(0.159)	(0.174)	(0.090)	(0.183)	(0.152)
	[0.259]	[0.458]	[0.475]	[0.511]	[0.563]	[0.583]	[0.240]
+ Prize	0.086**	-0.005	0.080	-0.052	0.207	0.184	0.204
(adding a prize)	(0.029)	(0.226)	(0.115)	(0.103)	(0.113)	(0.362)	(0.112)
	[0.030]	[0.985]	[0.566]	[0.629]	[0.160]	[0.687]	[0.202]
Group Size	0.049**	0.041	0.049**	0.049**	0.045*	0.048**	0.052**
	(0.020)	(0.033)	(0.019)	(0.019)	(0.020)	(0.020)	(0.019)
	[0.034]	[0.301]	[0.033]	[0.025]	[0.057]	[0.041]	[0.022]
Experience	0.136	0.131	0.110	0.142*	0.132*	0.137	0.139
	(0.071)	(0.069)	(0.156)	(0.070)	(0.067)	(0.072)	(0.075)
	[0.106]	[0.107]	[0.652]	[0.081]	[0.077]	[0.107]	[0.116]
Private	0.101	0.099	0.099	0.026	0.108*	0.103	0.111*
	(0.059)	(0.062)	(0.064)	(0.096)	(0.055)	(0.061)	(0.063)
	[0.115]	[0.139]	[0.142]	[0.859]	[0.069]	[0.116]	[0.086]
Men Share	0.039	0.032	0.042	0.044	-0.092	0.040	0.037
	(0.088)	(0.089)	(0.088)	(0.083)	(0.169)	(0.089)	(0.087)
	[0.677]	[0.746]	[0.656]	[0.606]	[0.655]	[0.673]	[0.679]
Median Age	-0.001	-0.001	-0.001	-0.001	-0.001	-0.002	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.001)	(0.003)
	[0.811]	[0.732]	[0.774]	[0.698]	[0.812]	[0.107]	[0.744]
German	-0.128	-0.132	-0.126	-0.119	-0.132	-0.130	-0.302**
	(0.080)	(0.081)	(0.082)	(0.087)	(0.085)	(0.081)	(0.073)
	[0.203]	[0.194]	[0.211]	[0.229]	[0.208]	[0.209]	[0.017]
+ Identity x Group Size		-0.049					
(making identity salient)		(0.075)					
		[0.544]					
+ Rank x Group Size		0.077					
(adding a ranking)		(0.065)					
		[0.349]					
+ Prize x Group Size		0.019					
(adding a prize)		(0.045)					
		[0.753]					
+ Identity x Experience			0.117				
(making identity salient)			(0.193)				
			[0.633]				
+ Rank x Experience			-0.112				
(adding a ranking)			(0.163)				
			[0.615]				
+ Prize x Experience			0.006				
(adding a prize)			(0.134)				
			[0.971]				

... continued on next page

Table 1.A.4: Team performance (completion, interactions) - continued

	Completed within 60 minutes						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
+ *Identity* x Private				0.185			
(making identity salient)				(0.184)			
				[0.552]			
+ *Rank* x Private				-0.161			
(adding a ranking)				(0.171)			
				[0.561]			
+ *Prize* x Private				0.158			
(adding a prize)				(0.135)			
				[0.287]			
+ *Identity* x Men Share					0.049		
(making identity salient)					(0.193)		
					[0.811]		
+ *Rank* x Men Share					0.319*		
(adding a ranking)					(0.124)		
					[0.059]		
+ *Prize* x Men Share					-0.256		
(adding a prize)					(0.200)		
					[0.377]		
+ *Identity* x Median Age						-0.002	
(making identity salient)						(0.004)	
						[0.851]	
+ *Rank* x Median Age						0.006	
(adding a ranking)						(0.006)	
						[0.482]	
+ *Prize* x Median Age						-0.003	
(adding a prize)						(0.011)	
						[0.823]	
+ *Identity* x German							0.690
(making identity salient)							(0.155)
							[0.236]
+ *Rank* x German							-0.318
(adding a ranking)							(0.167)
							[0.278]
+ *Prize* x German							-0.128
(adding a prize)							(0.133)
							[0.392]
Mean in Control	0.527	0.527	0.527	0.527	0.527	0.527	0.527
Observations	373	373	373	373	373	373	373
Team Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Staff FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weekday FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table displays coefficients from OLS regressions of whether a team completed the task within 60 minutes. All columns include room fixed effects. Each column indicates whether team controls (group size, share of males, experience, median age, language, private), staff, and weekday fixed effects are included. Standard errors in parentheses are clustered at the week level. *p*-values from score bootstrapping following Kline and Santos (2012) are listed in square brackets, with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

Table 1.A.5: Team performance (finishing times, interactions)

	Finishing time						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
+ *Identity*	1.590	-9.216	2.750	4.175	0.208	2.361	35.840
(making identity salient)	(1.117)	(6.773)	(2.683)	(2.505)	(2.713)	(3.358)	(3.740)
	[0.218]	[0.210]	[0.329]	[0.312]	[0.945]	[0.491]	[0.387]
+ *Rank*	-2.515**	12.942	-3.046	-8.517*	0.746	3.708	-34.370
(adding a ranking)	(0.836)	(8.091)	(3.011)	(2.315)	(2.129)	(3.039)	(3.554)
	[0.034]	[0.178]	[0.348]	[0.058]	[0.730]	[0.355]	[0.296]
+ *Prize*	-2.330*	-10.133	-4.148	6.096**	-3.601	-8.542	2.409
(adding a prize)	(1.319)	(6.405)	(3.395)	(1.560)	(2.351)	(6.488)	(2.264)
	[0.064]	[0.102]	[0.228]	[0.018]	[0.206]	[0.255]	[0.384]
Group Size	-1.408**	-1.811	-1.416**	-1.475**	-1.363**	-1.382**	-1.447**
	(0.508)	(0.771)	(0.495)	(0.491)	(0.519)	(0.520)	(0.501)
	[0.032]	[0.230]	[0.026]	[0.022]	[0.039]	[0.037]	[0.029]
Experience	-4.334**	-4.229**	-4.351	-4.590***	-4.256**	-4.282**	-4.437**
	(1.384)	(1.324)	(2.157)	(1.362)	(1.347)	(1.398)	(1.426)
	[0.011]	[0.012]	[0.186]	[0.010]	[0.011]	[0.011]	[0.011]
Private	-2.175*	-1.975	-1.992	-1.852	-2.203*	-2.316**	-2.003*
	(1.090)	(1.187)	(1.159)	(1.668)	(1.111)	(1.110)	(1.098)
	[0.067]	[0.115]	[0.117]	[0.334]	[0.070]	[0.046]	[0.077]
Men Share	-1.462	-1.256	-1.623	-1.474	-0.793	-1.503	-1.474
	(1.530)	(1.539)	(1.563)	(1.333)	(3.223)	(1.467)	(1.494)
	[0.325]	[0.403]	[0.286]	[0.270]	[0.821]	[0.302]	[0.316]
Median Age	0.060	0.073	0.063	0.076	0.058	0.120*	0.067
	(0.058)	(0.060)	(0.060)	(0.055)	(0.056)	(0.068)	(0.058)
	[0.250]	[0.151]	[0.243]	[0.105]	[0.263]	[0.077]	[0.194]
German	0.711	0.520	0.690	1.141	0.526	0.854	2.830
	(1.692)	(1.682)	(1.661)	(1.395)	(1.729)	(1.618)	(2.313)
	[0.702]	[0.767]	[0.708]	[0.513]	[0.773]	[0.645]	[0.356]
+ *Identity* x Group Size		2.451					
(making identity salient)		(1.417)					
		[0.122]					
+ *Rank* x Group Size		-3.399					
(adding a ranking)		(1.820)					
		[0.124]					
+ *Prize* x Group Size		1.649					
(adding a prize)		(1.313)					
		[0.238]					
+ *Identity* x Experience			-1.456				
(making identity salient)			(3.558)				
			[0.684]				
+ *Rank* x Experience			0.660				
(adding a ranking)			(3.416)				
			[0.898]				
+ *Prize* x Experience			2.513				
(adding a prize)			(3.411)				
			[0.422]				

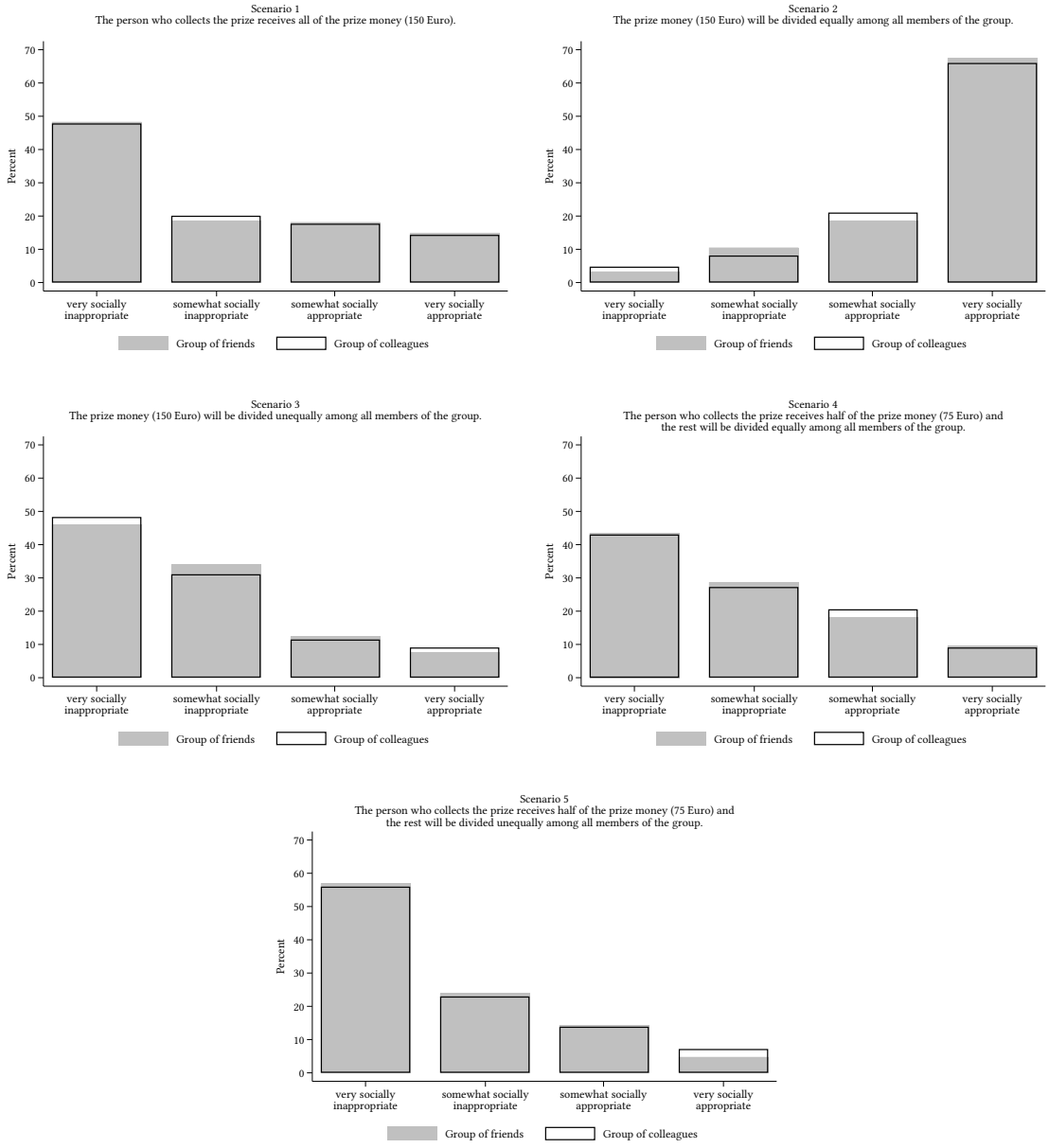
... continued on next page

Table 1.A.5: Team performance (finishing times, interactions) - continued

	Finishing time						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
+ *Identity* x Private				-2.759			
(making identity salient)				(2.673)			
				[0.472]			
+ *Rank* x Private				6.968*			
(adding a ranking)				(2.022)			
				[0.085]			
+ *Prize* x Private				-9.695***			
(adding a prize)				(2.260)			
				[0.005]			
+ *Identity* x Men Share					3.300		
(making identity salient)					(4.566)		
					[0.503]		
+ *Rank* x Men Share					-7.144		
(adding a ranking)					(3.515)		
					[0.120]		
+ *Prize* x Men Share					2.490		
(adding a prize)					(3.340)		
					[0.605]		
+ *Identity* x Median Age						-0.021	
(making identity salient)						(0.092)	
						[0.814]	
+ *Rank* x Median Age						-0.187	
(adding a ranking)						(0.101)	
						[0.218]	
+ *Prize* x Median Age						0.186	
(adding a prize)						(0.177)	
						[0.378]	
+ *Identity* x German							-34.494
(making identity salient)							(4.018)
							[0.389]
+ *Rank* x German							31.953
(adding a ranking)							(3.615)
							[0.323]
+ *Prize* x German							-4.954
(adding a prize)							(2.731)
							[0.205]
Mean in Control	56.470	56.470	56.470	56.470	56.470	56.470	56.470
Observations	373	373	373	373	373	373	373
Team Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Staff FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weekday FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table displays coefficients from Tobit regressions of finishing times. All columns include room fixed effects. Each column indicates whether team controls (group size, share of males, experience, median age, language, private), staff, and weekday fixed effects are included. Standard errors in parentheses are clustered at the week level. *p*-values from score bootstrapping following Kline and Santos (2012) are listed in square brackets, with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

Figure 1.A.4: Social norms of splitting a prize between friends and colleagues



Notes: The figure shows histograms of survey answers on the social appropriateness of splitting a monetary prize within a group of friends (taking part in their leisure time) or a group of colleagues (taking part in a team-building event). For each of the five scenarios, subjects had to evaluate the social appropriateness on a 4-point Likert-scale from "very socially inappropriate" to "very socially appropriate".

1.A.5 Willingness to explore original solutions and potential crowding out

Figure 1.A.5 illustrates the hint taking behavior over time and across treatments. In all treatments, teams request a similar number of hints. If anything, teams in *Prize* tend to take slightly fewer hints. OLS regressions on the number of hints (Table 1.A.6) confirm the non-parametric finding that neither component, *Identity*, *Rank* nor *Prize* affect the willingness to explore original solutions, also when controlling for team characteristics, adding staff, or weekday fixed effects. In fact, all coefficients are small in magnitude, sometimes switch to the opposite sign, and are far from statistically significant.

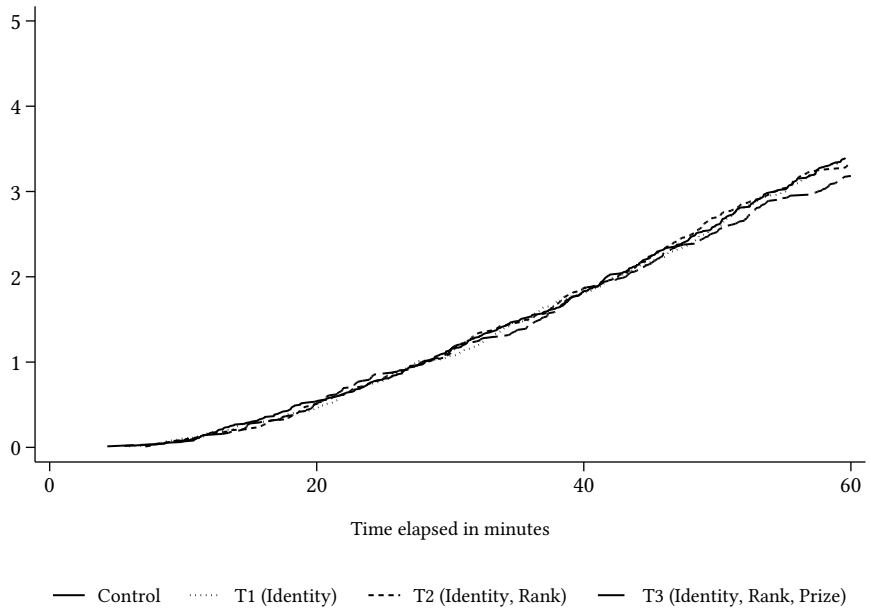
Even though the willingness to explore new and original solutions does not seem to be crowded out if measured by the total number of hints requested, it would still be conceivable that teams request their hints earlier. This would effectively also allow them to rely on external help early on and thus arrive at the solution quicker. Table 1.A.7 shows the coefficients of Tobit regressions on the timing of hints using treatment components as explanatory variables.[21] The results are again small in magnitude and indistinguishable from zero. The step-wise introduction of additional controls and fixed effects does not affect this result.

To shed light on whether particularly (un)successful teams differ in their willingness to explore original solutions, we also present results from linear regressions within quantiles (based on residualized finishing times) in Figure 1.A.6. Panel A shows the difference in the number of hints taken in *Identity* as compared to *Control*. Panel B compares adding *Rank* to only having component *Identity*. Panel C compares the addition of *Prize* on top of *Rank*. Panel D provides the comparison between *Control* and *T3 (Identity, Rank, Prize)*. No clear and consistent picture emerges: none of the components seem to affect the number of hints taken across the entire performance spectrum.

To analyze whether our treatments reduced a team's intrinsic motivation, Table 1.A.8 presents results from Probit regressions on the marginal effects of the *Identity*, *Rank*, and *Prize* components on purchasing a voucher. As in previous analyses, we add additional controls and fixed effects in each column. The results speak clearly against any crowding out of intrinsic motivation for future participation.

[21] We assigned a time of 60 minutes for all unused hints.

Figure 1.A.5: Hint taking over time



Notes: The figure shows the cumulative distribution of hints by minute in *Control*, *T1 (Identity)*, *T2 (Identity, Rank)*, and *T3 (Identity, Rank, Prize)*.

Table 1.A.6: Willingness to explore original solutions (number of hints)

	Number of hints			
	(1)	(2)	(3)	(4)
+ Identity	-0.058	-0.038	0.048	0.075
(making identity salient)	(0.308)	(0.297)	(0.299)	(0.300)
	[0.868]	[0.916]	[0.901]	[0.845]
+ Rank	0.028	0.040	0.102	0.100
(adding a ranking)	(0.342)	(0.305)	(0.296)	(0.288)
	[0.924]	[0.895]	[0.821]	[0.835]
+ Prize	-0.142	-0.171	-0.214	-0.213
(adding a prize)	(0.279)	(0.248)	(0.211)	(0.187)
	[0.642]	[0.530]	[0.415]	[0.398]
Mean in Control	3.393	3.393	3.393	3.393
Observations	373	373	373	373
Team Controls	No	Yes	Yes	Yes
Staff FE	No	No	Yes	Yes
Weekday FE	No	No	No	Yes

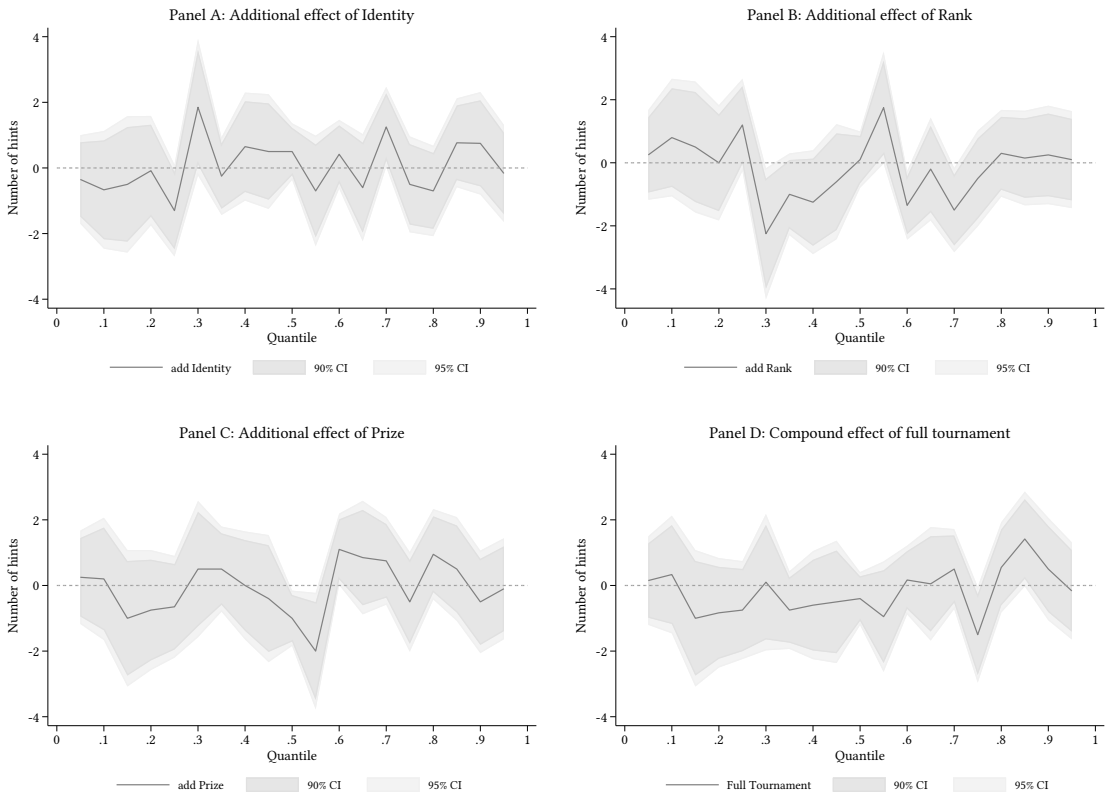
Notes: The table displays coefficients from OLS regressions of number of hints. The main explanatory variables are indicators whether the observation stems from a treatment that included the component(s) *Identity*, *Rank*, or *Prize*. All columns include room fixed effects. Each column indicates whether team controls (group size, share of males, experience, median age, language, private), staff, and weekday fixed effects are included. Standard errors in parentheses are clustered at the week level. p-values from score bootstrapping following Kline and Santos (2012) are listed in square brackets, with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

Table 1.A.7: Willingness to explore original solutions (timing of hints)

	Timing of hints				
	1st hint	2nd hint	3rd hint	4th hint	5th hint
	(1)	(2)	(3)	(4)	(5)
+ Identity	1.279	-0.358	0.105	-0.034	-1.636
(making identity salient)	(1.259)	(1.899)	(2.199)	(2.238)	(2.241)
	[0.368]	[0.866]	[0.968]	[0.985]	[0.527]
+ Rank	-2.306	-1.977	-1.524	-2.573	0.421
(adding a ranking)	(1.674)	(2.317)	(2.436)	(2.814)	(2.605)
	[0.203]	[0.447]	[0.567]	[0.376]	[0.881]
+ Prize	-0.155	1.808	2.960	3.504	2.619
(adding a prize)	(1.829)	(2.108)	(2.184)	(2.402)	(1.999)
	[0.965]	[0.454]	[0.225]	[0.250]	[0.301]
Mean in Control	22.990	37.243	47.715	55.072	58.448
Observations	373	373	373	373	373
Team Controls	Yes	Yes	Yes	Yes	Yes
Staff FE	Yes	Yes	Yes	Yes	Yes
Weekday FE	Yes	Yes	Yes	Yes	Yes

Notes: The table displays coefficients from Tobit regressions of timing of hints. The main explanatory variables are indicators whether the observation stems from a treatment that included the component(s) *Identity*, *Rank*, or *Prize*. All columns include room fixed effects. Each column indicates whether team controls (group size, share of males, experience, median age, language, private), staff, and weekday fixed effects are included. Standard errors in parentheses are clustered at the week level. *p*-values from score bootstrapping following Kline and Santos (2012) are listed in square brackets, with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

Figure 1.A.6: OLS regressions on number of hints (within quantiles)



Notes: The figure shows OLS regressions (within quantiles sorted by residualized finishing time) on number of hints. Panel A shows the additional effect of salient team identity. Panel B shows the additional effect of a public ranking. Panel C shows the additional effect of a monetary prize. And Panel D shows the overall effect of a tournament with a monetary prize (compares *T3 (Identity, Rank, Prize)* to *Control*). The line at zero marks the number of hints in the comparison group. Negative (positive) values indicate reductions (increases) in the number of hints due to *Identity* (Panel A), *Rank* (Panel B), *Prize* (Panel C), or due to adding all tournament features simultaneously (Panel D).

Table 1.A.8: Purchased a voucher

	Purchased a voucher			
	(1)	(2)	(3)	(4)
+ Identity	0.012	-0.004	0.005	0.009
(making identity salient)	(0.025)	(0.028)	(0.024)	(0.027)
	[0.575]	[0.906]	[0.831]	[0.742]
+ Rank	0.053	0.041	0.019	0.014
(adding a ranking)	(0.038)	(0.033)	(0.028)	(0.026)
	[0.276]	[0.332]	[0.551]	[0.621]
+ Prize	-0.009	0.004	0.001	0.010
(adding a prize)	(0.057)	(0.048)	(0.042)	(0.042)
	[0.890]	[0.928]	[0.984]	[0.794]
Mean in Control	0.179	0.179	0.179	0.179
Observations	373	373	373	373
Team Controls	No	Yes	Yes	Yes
Staff FE	No	No	Yes	Yes
Weekday FE	No	No	No	Yes

Notes: The table displays average marginal effects from Probit regressions of whether a team purchased a voucher. The main explanatory variables are indicators whether the observation stems from a treatment that included the component(s) *Identity*, *Rank*, or *Prize*. All columns include room fixed effects. Each column indicates whether team controls (group size, share of males, experience, median age, language, private), staff, and weekday fixed effects are included. Standard errors in parentheses are clustered at the week level. *p*-values from score bootstrapping following Kline and Santos (2012) are listed in square brackets, with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

1.A.6 Water damage

For our main data analysis, we removed five observations because of water damage to ETR's equipment resulting from a burst pipe. Table 1.A.9 repeats the specifications from Table 1.2 but includes the five omitted data points. The results are very similar.

Table 1.A.9: Team performance (including observations affected by water damage)

	Completed within 60 minutes				Finishing time			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
+ Identity	0.085	-0.097	-0.048	-0.044	1.378	1.910	1.668	1.590
(making identity salient)	(0.051)	(0.065)	(0.059)	(0.055)	(0.870)	(0.970)	(1.180)	(1.117)
	[0.198]	[0.240]	[0.434]	[0.446]	[0.220]	[0.146]	[0.205]	[0.217]
+ Rank	0.103	0.092	0.080	0.078	-2.789*	-2.583*	-2.575**	-2.515**
(adding a ranking)	(0.045)	(0.048)	(0.048)	(0.044)	(0.856)	(0.801)	(0.851)	(0.836)
	[0.127]	[0.181]	[0.229]	[0.184]	[0.057]	[0.051]	[0.034]	[0.034]
+ Prize	0.090**	0.091**	0.078**	0.083**	-2.214**	-2.391**	-2.200**	-2.330*
(adding a prize)	(0.031)	(0.028)	(0.030)	(0.026)	(1.047)	(1.224)	(1.275)	(1.319)
	[0.047]	[0.032]	[0.032]	[0.020]	[0.042]	[0.033]	[0.040]	[0.064]
Mean in Control	0.527	0.527	0.527	0.527	56.470	56.470	56.470	56.470
Observations	378	378	378	378	378	378	378	378
Team Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Staff FE	No	No	Yes	Yes	No	No	Yes	Yes
Weekday 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)). The main explanatory variables are indicators whether the observation stems from a treatment that included the component(s) *Identity*, *Rank*, or *Prize*. All columns include room fixed effects. Each column indicates whether team controls (group size, share of males, experience, median age, language, private), staff, and weekday fixed effects are included. Standard errors in parentheses are clustered at the week level. *p*-values from score bootstrapping following Kline and Santos (2012) are listed in square brackets, with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

Appendix to Chapter 2

2.A.1 Additional robustness analyses

In this section, we present results on the robustness of the observed treatment effect. Table 2.A.1 repeats the specifications from Table 2.2 but excludes the 12 observations, where ETR staff implemented the wrong treatment. The results are very similar. All coefficients are of similar magnitude and only one specification lacks statistical significance at conventional levels (Column (3)). Table 2.A.2 reports findings from a linear probability model estimating the impact of *Leadership* on the probability to solve the task and generalized linear model estimations on teams' finishing times.

Table 2.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)–(4)), and Tobit regressions of finishing time (Columns (5)–(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 2.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

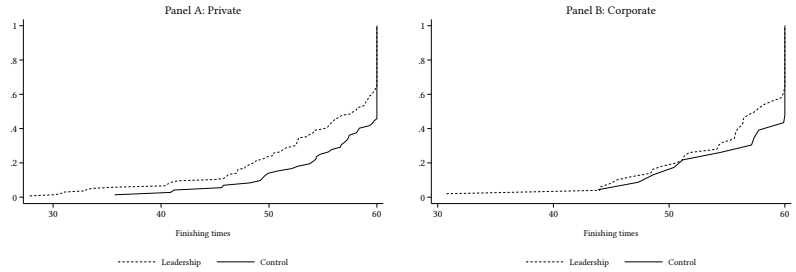
Notes: The table displays coefficients from OLS regressions of whether a team completed the task within 60 minutes (Columns (1)–(4)) and GLM regressions (with log link) of finishing time (Columns (5)–(8)) on our *Leadership* indicator (with *Control* as base category). All columns include room fixed 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 *** = $p < 0.01$.

2.A.2 Heterogeneity in reactions to *Leadership*

In this section, we briefly investigate heterogeneous reactions to treatments (see Tables 2.A.3 and 2.A.4). We do not find strong interactions of our *Leadership* condition and observable team characteristics such as group size, experience, 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 5% 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 because 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 further illustrate potential differences between these groups, we present separate cumulative distributions of finishing times in Appendix Figure 2.A.1 in addition to the regression results shown in Appendix Tables 2.A.3 and 2.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 2.A.3 and 2.A.4, Column (4)).

Figure 2.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 2.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 2.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 *** = $p < 0.01$.

2.A.3 Team characteristics and choosing a leader

Table 2.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 chosen leader's gender. 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 do not speak German.

Table 2.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, Column (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 *** = $p < 0.01$.

2.A.4 Results from customer survey

To analyze how teams perceived their experience and performance, Table 2.A.6 presents the results from OLS regressions as well as the second stage from 2SLS regressions following the approach recommended in Angrist and Pischke (2008, p.142).[22] 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 2.A.6: Customer survey

	Value for Money (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	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). The survey included the following questions: "Are you satisfied with the price-performance ratio?" (Value for Money), "How did you like the experience in general?" (Satisfaction), "How hard did you try?" (Effort), "How much were you motivated as a team?" (Motivation), and "How well were you organized as a team?" (Coordination). Participants evaluated these questions on an eight-point Likert scale (ranging from 1="not at all" to 8="very much"). All variables are standardized with mean zero and a standard deviation of one. For 2SLS (Panel B), we follow the procedure outlined by Angrist and Pischke (2008): we first 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. We then 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 *** = $p < 0.01$.

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

Appendix to Chapter 3

3.A.1 Perceptions

Table 3.A.1 shows the results of the conditional logit model discussed in Section 3.3.1. Specifications (1) and (5) reflect the specifications presented in Figure 3.3. Specifications (2) – (4) show specifications in which we additionally study whether the perceived relative importance of governance structures depends on the team composition (i.e., group size, diversity, and experience) for escape challenges. We find that the perceived efficacy of Bonus incentives does not interact with team composition. Rank incentives are perceived as relatively less effective for larger, more diverse, and experienced groups. Tournaments with prices are perceived as relatively less effective for experienced groups and diverse teams are perceived to thrive even more with female leadership. The perceived efficacy of male leadership does not depend on team characteristics.

Specifications (6) – (8) show specifications in which we additionally study whether the perceived relative importance of governance structures depends on the team composition (i.e., group size, diversity, and experience) for the web developer task. We find that the perceived efficacy of Bonus incentives also does not interact with team composition in the Web developer task and that Rank incentives are also perceived as relatively less effective for larger teams (but are not perceived to be significantly less effective for more diverse or experienced groups). Interestingly, in the web development task, diverse teams are perceived to thrive even more with Rank incentives, and - akin to the escape challenge - with female leadership. The perceived efficacy of male leadership does not depend statistically significantly on team size or gender diversity of teams, but male leadership is expected to be more effective for teams with at least one experienced member.

Table 3.A.1: Perceptions (HR experts)

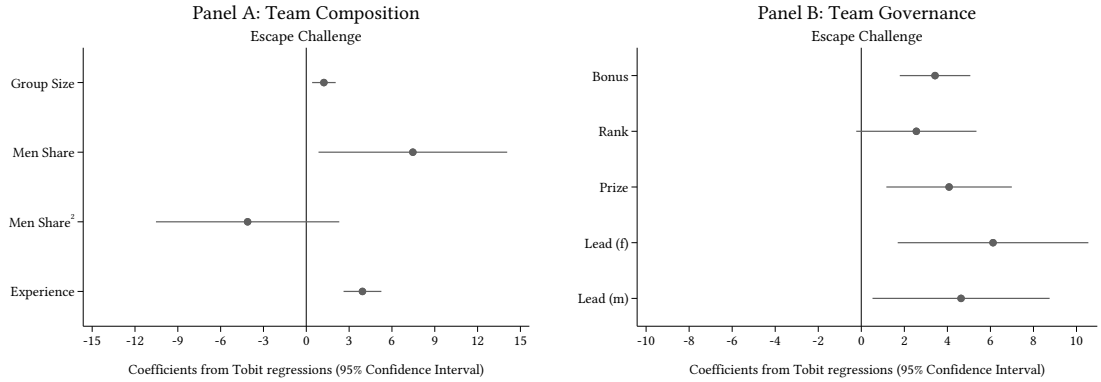
	Conditional logit: Choice							
	Escape Challenge				Web Development			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bonus	0.811***	0.753***	0.836***	0.892***	0.874***	0.662***	0.844***	0.865***
	(0.030)	(0.247)	(0.069)	(0.058)	(0.027)	(0.209)	(0.055)	(0.045)
Rank	0.903***	1.309***	1.044***	1.005***	0.978***	1.379***	1.005***	1.003***
	(0.033)	(0.227)	(0.070)	(0.055)	(0.028)	(0.210)	(0.060)	(0.052)
Prize	0.953***	1.048***	1.027***	1.041***	1.043***	1.210***	0.994***	1.056***
	(0.034)	(0.283)	(0.080)	(0.058)	(0.029)	(0.258)	(0.065)	(0.047)
Lead (f)	0.636***	0.339	0.575***	0.658***	0.723***	1.006***	0.616***	0.713***
	(0.036)	(0.309)	(0.071)	(0.066)	(0.030)	(0.221)	(0.068)	(0.054)
Lead (m)	0.593***	0.691***	0.553***	0.629***	0.682***	1.183***	0.582***	0.597***
	(0.040)	(0.266)	(0.153)	(0.066)	(0.036)	(0.320)	(0.172)	(0.061)
Group Size	0.122***	0.128***	0.120***	0.118***	0.122***	0.155***	0.120***	0.122***
	(0.012)	(0.030)	(0.011)	(0.012)	(0.011)	(0.023)	(0.011)	(0.011)
Men Share	0.541***	0.548***	0.586***	0.547***	0.502***	0.520***	0.183	0.491***
	(0.102)	(0.095)	(0.210)	(0.103)	(0.093)	(0.088)	(0.177)	(0.091)
Men Share^2	-0.524***	-0.537***	-0.518**	-0.531***	-0.565***	-0.586***	-0.225	-0.554***
	(0.102)	(0.098)	(0.206)	(0.103)	(0.097)	(0.091)	(0.185)	(0.095)
Experience	1.586***	1.582***	1.588***	1.676***	1.639***	1.638***	1.637***	1.630***
	(0.017)	(0.017)	(0.017)	(0.056)	(0.016)	(0.016)	(0.016)	(0.044)
Group Size x Bonus		0.012				0.042		
		(0.047)				(0.040)		
Men Share x Bonus			-0.009				0.389	
			(0.341)				(0.285)	
Men Share^2 x Bonus			-0.011				-0.430	
			(0.331)				(0.295)	
Experience x Bonus				-0.125				0.016
				(0.087)				(0.076)
Group Size x Rank		-0.077*				-0.077*		
		(0.043)				(0.040)		
Men Share x Rank			-0.902***				-0.002	
			(0.327)				(0.285)	
Men Share^2 x Rank			0.835***				-0.063	
			(0.323)				(0.281)	
Experience x Rank				-0.186**				-0.074
				(0.081)				(0.074)
Group Size x Prize		-0.016				-0.033		
		(0.053)				(0.048)		
Men Share x Prize			-0.100				0.652**	
			(0.391)				(0.324)	
Men Share^2 x Prize			-0.042				-0.713**	
			(0.377)				(0.325)	
Experience x Prize				-0.161*				-0.052
				(0.088)				(0.076)
Group Size x Lead (f)		0.057				-0.053		
		(0.060)				(0.043)		
Men Share x Lead (f)			0.699*				0.888**	
			(0.407)				(0.385)	
Men Share^2 x Lead (f)			-0.894*				-0.993**	
			(0.501)				(0.444)	
Experience x Lead (f)				-0.011				0.016
				(0.096)				(0.079)
Group Size x Lead (m)		-0.017				-0.094		
		(0.052)				(0.059)		
Men Share x Lead (m)			0.354				0.460	
			(0.642)				(0.692)	
Men Share^2 x Lead (m)			-0.410				-0.379	
			(0.540)				(0.588)	
Experience x Lead (m)				-0.051				0.152*
				(0.094)				(0.091)
Observations	78,000	78,000	78,000	78,000	78,000	78,000	78,000	78,000

Notes: The table displays results from a Conditional logit choice model. The regressions include team governance indicators (with Control as base alternative), team composition (group size, share of males, experience) and interaction terms (team governance and team composition). Standard errors are clustered on the choice set level (with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$).

3.A.2 Misperceptions

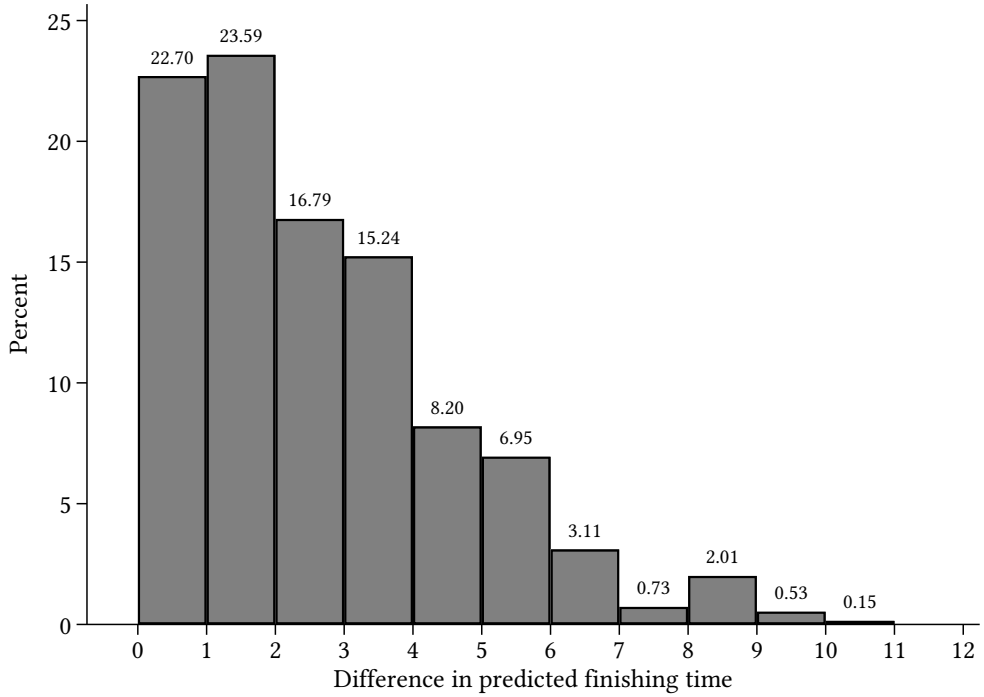
Figure 3.A.1 shows coefficient plots from Tobit regressions of actual performance data on team composition and team governance structures. Figure 3.A.2 illustrates the differences in predicted finishing times for those comparisons, where the HR experts did not choose the 'Naïve Expert' option. Table 3.A.2 provides results from a pairwise comparison of differences of perceptions about team governance structures between the 'Naïve Expert' and the HR experts.

Figure 3.A.1: Coefficients from actual performance data



Notes: The figure shows coefficient plots from Tobit regressions (with 95% confidence bands). Panel A (Panel B) shows coefficients of team performance on team composition (team governance structures). A positive (negative) value indicates positive (negative) effects on team performance.

Figure 3.A.2: Winning margin for 'incorrect' choices



Notes: The figure shows the differences in predicted finishing times for those comparisons, where the HR experts did not choose the 'Naïve Expert' option.

Table 3.A.2: Comparison between 'Naïve Expert' and HR experts

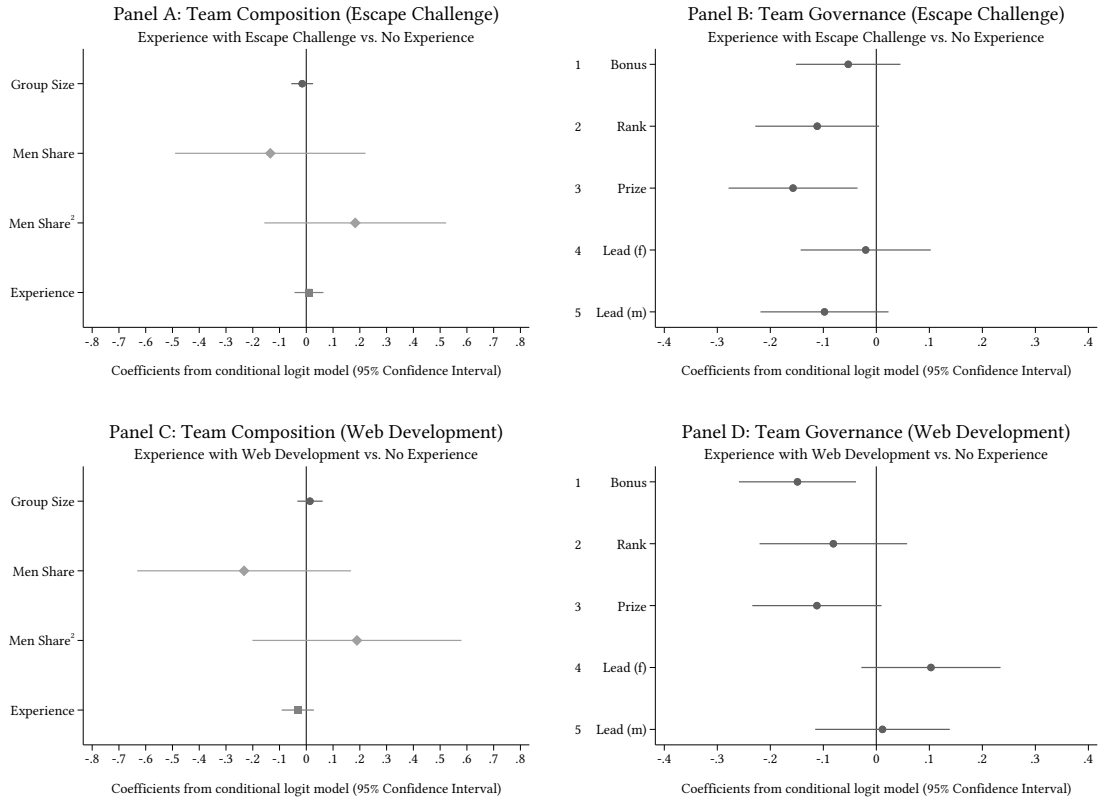
	Conditional logit (McFadden's): Choice					
	Control	Bonus	Rank	Prize	Lead (f)	Lead (m)
Constant	base	0.812***	0.905***	0.955***	0.637***	0.594***
	alternative	(0.030)	(0.033)	(0.034)	(0.035)	(0.040)
Naïve Expert		0.545*	-0.039	0.479	1.430***	0.780**
		(0.317)	(0.329)	(0.355)	(0.356)	(0.329)
Constant		base	0.092***	0.142***	-0.175***	-0.219***
		alternative	(0.033)	(0.035)	(0.038)	(0.040)
Naïve Expert			-0.584*	-0.067	0.884**	0.234
			(0.310)	(0.355)	(0.344)	(0.296)
Constant			base	0.050	-0.267***	-0.311***
			alternative	(0.033)	(0.039)	(0.040)
Naïve Expert				0.517	1.469***	0.818***
				(0.328)	(0.341)	(0.307)
Constant				base	-0.317***	-0.361***
				alternative	(0.040)	(0.044)
Naïve Expert					0.951***	0.301
					(0.349)	(0.315)
Constant					base	-0.044
					alternative	(0.047)
Naïve Expert						-0.650**
						(0.330)
Observations	78,390	78,390	78,390	78,390	78,390	78,390

Notes: The table displays results from a Conditional logit (McFadden's) choice model (base alternative as labelled). Constant indicates whether HR experts are more or less likely to pick another alternative. Naïve Expert indicates differences to the predictions using our field data. Standard errors in parentheses are clustered on the choice set level, with * = $p < 0.10$, ** = $p < 0.05$ and *** = $p < 0.01$.

3.A.3 Discussion

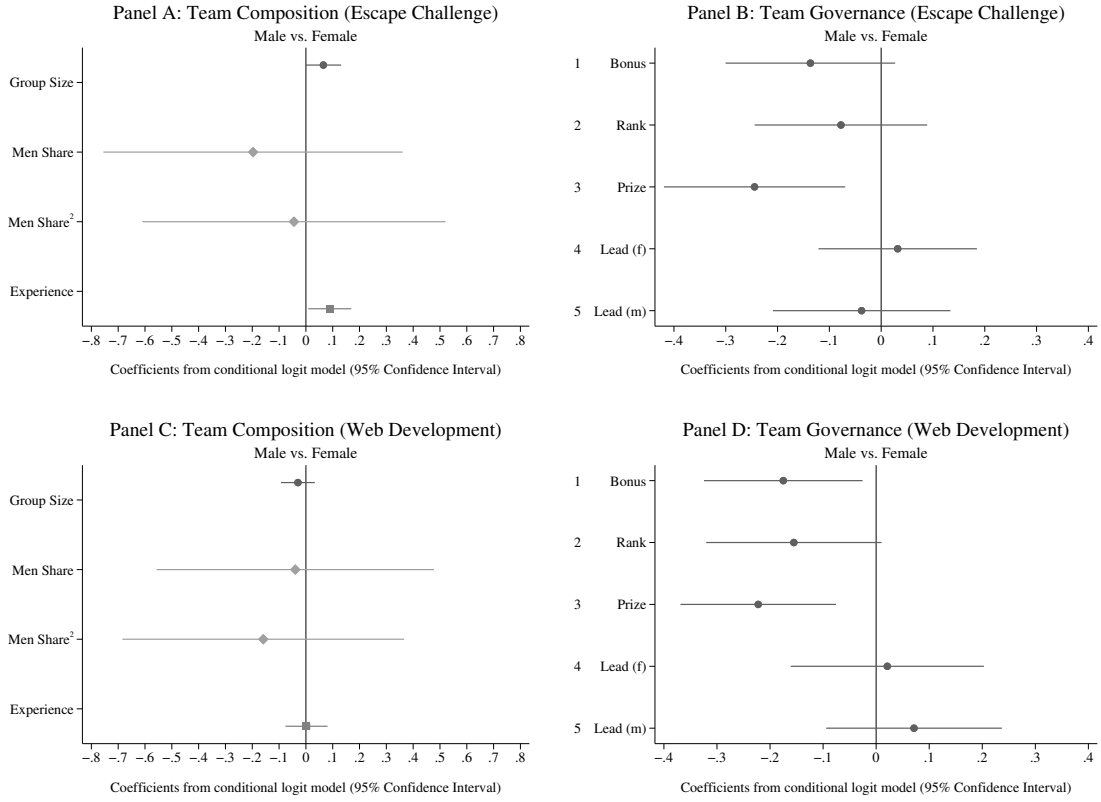
Figure 3.A.3 illustrates differences in perceptions with respect to team composition and team governance structures for experienced and non-experienced HR experts. Figure 3.A.4 illustrates differences in second-order beliefs with respect to team composition and team governance structures for males and females.

Figure 3.A.3: Comparison between experienced and non-experienced HR experts



Notes: The figure shows coefficient plots from conditional logit models (with 95% confidence bands). Panel A and Panel C show differences in perceptions with respect to team composition for experienced and non-experienced HR experts. Panel B and Panel D show differences in perceptions with respect to team governance structures. A positive (negative) value indicates that experienced HR experts expect this factor to be more (less) important than non-experienced HR experts.

Figure 3.A.4: Comparison between males and females (second-order beliefs)



Notes: The figure shows coefficient plots from conditional logit models (with 95% confidence bands). Panel A and Panel C show differences in second-order beliefs with respect to team composition for males and females. Panel B and Panel D show differences in second-order beliefs with respect to team governance structures. A positive (negative) value indicates that males expect that other experts perceive this factor to be more (less) important than females do.

Appendix to Chapter 4

4.A.1 Expert survey

Table 4.A.1 displays expected effect sizes of task assignment (i.e. which task the worker is working on) on performance and satisfaction. Thereby, I analyze differences in the probability of a decrease and an increase in performance and satisfaction, as well as differences in overall expected changes of these outcome variables. Table 4.A.2 displays expected effect sizes of different task assignment procedures (i.e. how tasks are assigned) on performance. Table 4.A.3 displays expected effect sizes of different task assignment procedures (i.e. how tasks are assigned) on satisfaction. Figure 4.A.1 represents word clouds of the frequency of used words in the open text question regarding task assignment in practice.

Table 4.A.1: Expected effect on performance and satisfaction (task assignment)

	Performance			Satisfaction		
	Decrease (1)	Increase (2)	Change (3)	Decrease (4)	Increase (5)	Change (6)
Less productive task	-0.079***	0.087***	6.024*	-0.078***	0.079***	7.085*
	(0.029)	(0.030)	(3.507)	(0.029)	(0.030)	(3.908)
Preferred task	-0.385***	0.382***	29.864***	-0.435***	0.426***	42.108***
	(0.031)	(0.031)	(2.790)	(0.028)	(0.028)	(3.361)
Productive task	-0.311***	0.326***	21.654***	-0.369***	0.364***	32.817***
	(0.027)	(0.028)	(3.208)	(0.027)	(0.027)	(3.553)
Worker is female	-0.058**	0.053**	7.227***	-0.001	-0.002	1.735
	(0.023)	(0.024)	(1.990)	(0.023)	(0.023)	(2.392)
Less preferred task (mean)	0.499	0.480	7.553	0.563	0.434	0.208
Less productive = Preferred	0.000	0.000	0.000	0.000	0.000	0.000
Less productive = Productive	0.000	0.000	0.000	0.000	0.000	0.000
Preferred = Productive	0.022	0.086	0.003	0.045	0.062	0.002
Observations	3,200	3,200	3,200	3,200	3,200	3,200
Characteristics	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table displays expected effect sizes on performance (Columns (1)–(3)) and satisfaction (Columns (4)–(6)). Columns (1) and (4) (Columns (2) and (5)) display the average marginal effects from Probit regressions of whether practitioners expect a decrease (an increase) of performance and satisfaction on the task a worker is working on (*Less preferred* task as the base category). Columns (3) and (6) display coefficients from OLS regressions of the expected change of performance and satisfaction (in percent) on the task a worker is working on (*Less preferred* task as the base category). All columns include participant characteristics (gender, age, educational background, whether they work full- or part time). Standard errors in parentheses are clustered at the subject level, with significance levels * = $p < 0.10$, ** = $p < 0.05$, and *** = $p < 0.01$.

Table 4.A.2: Expected effect on performance (assignment procedure)

	Performance				
	All tasks	Not preferred	Preferred	Not productive	Productive
	(1)	(2)	(3)	(4)	(5)
Productivity	-1.588	-3.104	0.535	-1.475	-2.318
	(1.407)	(2.740)	(2.580)	(2.970)	(2.556)
Preferences	-1.046	-1.515	-0.460	2.399	-4.611*
	(1.320)	(2.031)	(2.132)	(2.939)	(2.635)
Self	0.976	5.644**	-0.812	0.076	-1.061
	(1.365)	(2.540)	(2.604)	(2.251)	(2.943)
Worker is female	7.064***	2.772	3.628	11.906**	11.239***
	(2.323)	(4.832)	(3.634)	(4.848)	(3.925)
Mean in Random	22.84	7.30	37.67	14.42	32.01
Productivity = Preferences	0.691	0.538	0.671	0.157	0.467
Productivity = Self	0.053	0.000	0.571	0.593	0.665
Preferences = Self	0.135	0.008	0.882	0.419	0.238
Observations	3,200	808	808	792	792
Characteristics	Yes	Yes	Yes	Yes	Yes

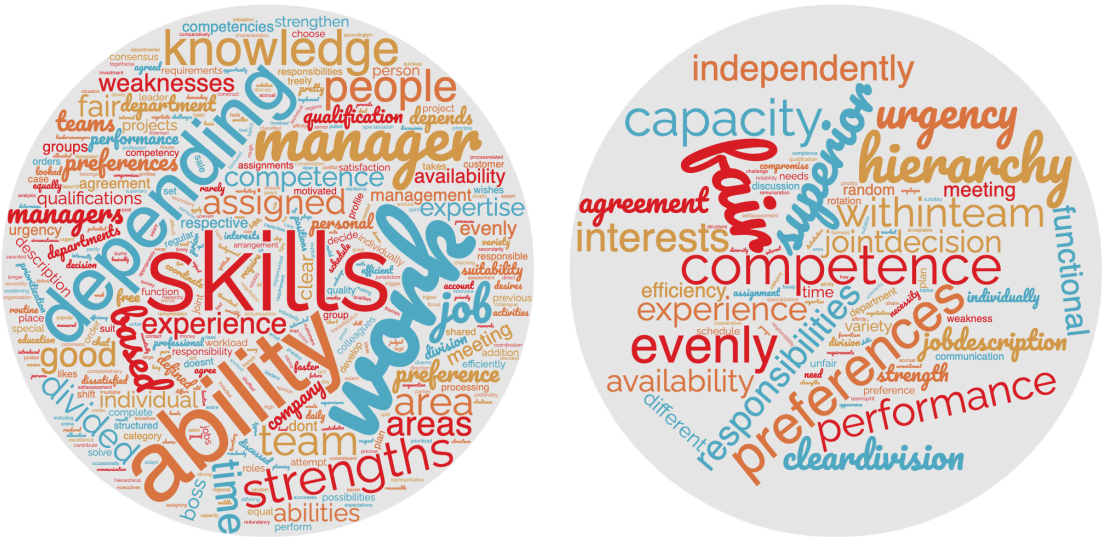
Notes: The table displays coefficients from OLS regressions of the expected change of performance (in percent) on the task assignment procedure (with treatment *Random* as base category). While Column (1) includes all tasks, Column (2)–Column (5) differentiate between advantaged and disadvantaged workers. All columns include participant characteristics (gender, age, educational background, whether they work full- or part time). Standard errors in parentheses are clustered at the subject level, with significance levels * = $p < 0.10$, ** = $p < 0.05$, and *** = $p < 0.01$.

Table 4.A.3: Expected effect on satisfaction (assignment procedure)

	Satisfaction				
	All tasks	Not preferred	Preferred	Not productive	Productive
	(1)	(2)	(3)	(4)	(5)
Productivity	-0.661	-0.342	0.500	-1.949	-0.884
	(1.379)	(2.757)	(2.495)	(3.147)	(2.695)
Preferences	0.174	1.658	2.381	-0.192	-3.227
	(1.427)	(2.700)	(2.343)	(2.871)	(2.961)
Self	2.624*	7.455**	1.777	1.283	-0.101
	(1.495)	(2.984)	(2.830)	(3.024)	(2.828)
Worker is female	1.431	0.351	1.529	6.615	-1.217
	(2.963)	(5.289)	(3.863)	(5.439)	(4.359)
Mean in Random	20.43	-1.99	41.17	8.03	34.53
Productivity = Preferences	0.534	0.470	0.386	0.586	0.451
Productivity = Self	0.021	0.007	0.601	0.378	0.774
Preferences = Self	0.080	0.065	0.789	0.650	0.272
Observations	3,200	808	808	792	792
Characteristics	Yes	Yes	Yes	Yes	Yes

Notes: The table displays coefficients from OLS regressions of the expected change of satisfaction (in percent) on the task assignment procedure (with treatment *Random* as base category). While Column (1) includes all tasks, Column (2)–Column (5) differentiate between advantaged and disadvantaged workers. All columns include participant characteristics (gender, age, educational background, whether they work full- or part time). Standard errors in parentheses are clustered at the subject level, with significance levels * = $p < 0.10$, ** = $p < 0.05$, and *** = $p < 0.01$.

Figure 4.A.1: Task assignment procedures in practice (open text)



Notes: The figure represents the frequency of how often a certain word was used in the open text form regarding task assignment in practice (only taking reasonable answers into account). The left figure shows the frequency of words for all answers only excluding stop words (task and employee were added as stop words). The right figure shows the frequency of words for a short summary of each statement (e.g. competence, ability, knowledge summarized as competence).

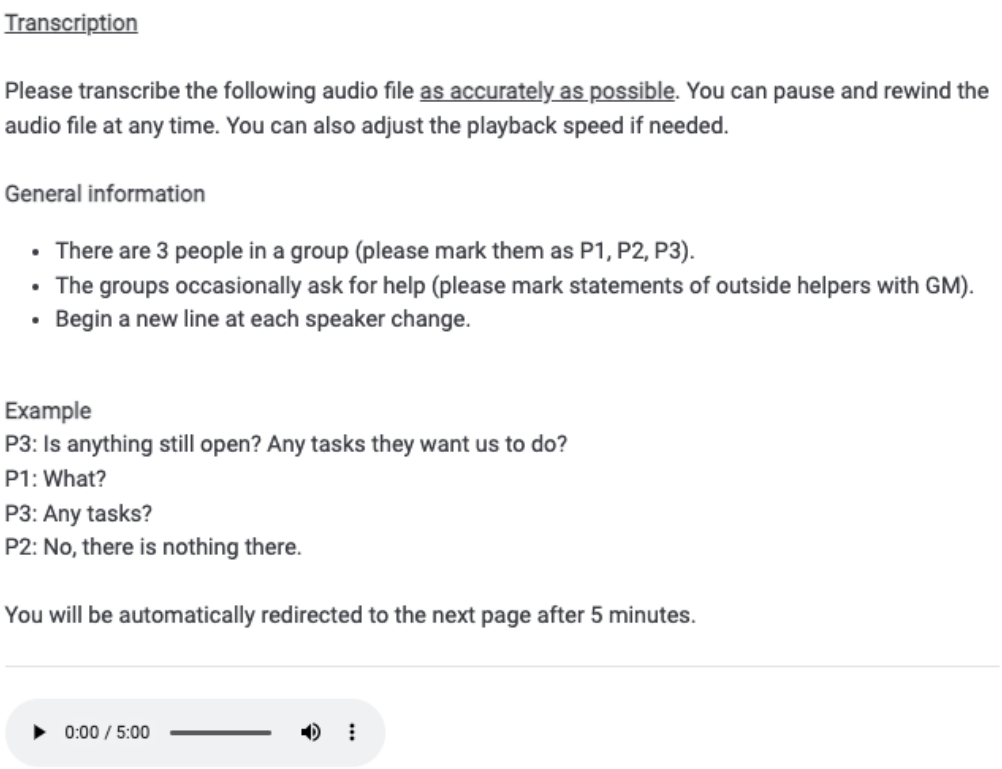
4.A.2 Treatment implementation and instructions

I implement five experimental variations, which I randomize on a daily level to avoid treatment spillovers. The respective experimental variation will be implemented right after the first evaluation phase. After clarifying any open questions, treatments are implemented according to the following script:

- *"Both tasks are equally important for our analyses, and therefore some people will transcribe audio data, while other people will evaluate transcribed files. Who is working on which task is determined..."*
- **Control**: *"... randomly. Your tablets are either marked with headphones or with a magnifying glass. Thus, both of you [S1 & S2] work on the transcription task (the analysis task)."*
- **Treatment**:
  - **Random**: *"... randomly. Your tablets are either marked with headphones or with a magnifying glass. Thus, you [S1] work on the transcription task, and you [S2] work on the analysis task (or vice versa)."*
  - **Productivity**: *"... by your expected productivity. You [S1] expect to perform better in the transcription task, and you [S2] expect to perform better in the analysis task. / Both of you [S1 & S2] expect to perform better in the transcription task. But in relative terms, you [S1] expect to perform better in the transcription task, and you [S2] expect to perform better in the analysis task. Thus, you [S1] work on the transcription task, and you [S2] work on the analysis task (or vice versa)."*
  - **Preferences**: *"... by your preferences. You [S1] prefer the transcription task, and you [S2] prefer the analysis task. / Both of you [S1 & S2] prefer the transcription task. But in relative terms, you [S1] prefer the transcription task more, and you [S2] prefer the analysis task more. Thus, you [S1] work on the transcription task, and you [S2] work on the analysis task (or vice versa)."*
  - **Self**: *"... by yourself. You now have some time to briefly discuss and jointly find an agreement. [Discussion]. Thus, you [S1] work on the transcription task, and you [S2] work on the analysis task (or vice versa)."*

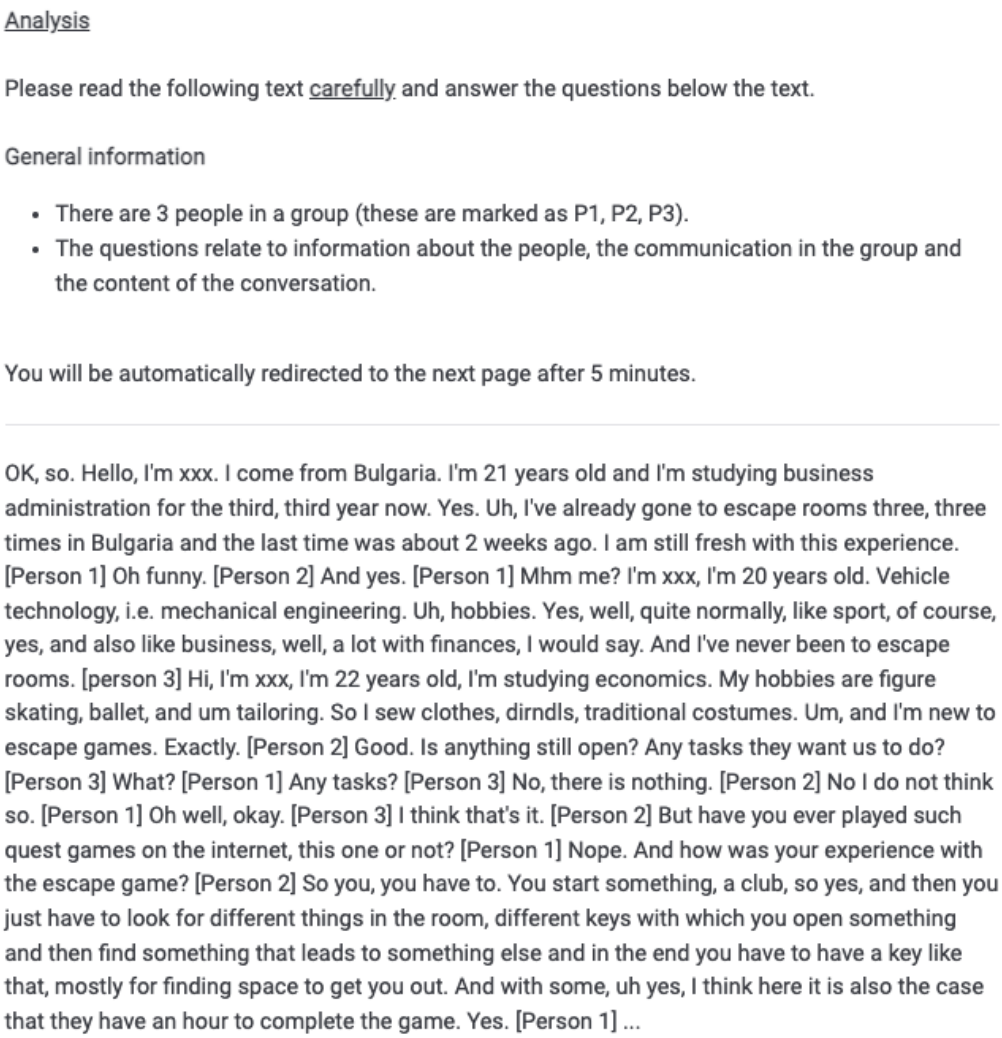
The figures below illustrate the instructions for the transcription task (Figure 4.A.2) and the analysis task (Figure 4.A.3), as well as the evaluation dimensions for the analysis task (Figure 4.A.4).

Figure 4.A.2: Instructions for the transcription task



Notes: The figure illustrates instructions for the transcription task (translated from German).

Figure 4.A.3: Instructions for the analysis task



Notes: The figure illustrates instructions for the analysis task (translated from German).

Figure 4.A.4: Evaluation dimensions for the analysis task

Notes: The figure illustrates the evaluation dimensions for the analysis task (translated from German).

Bibliography

Ai, W., Chen, Y., Mei, Q., Ye, J., and Zhang, L. (2022). Putting teams into the gig economy: A field experiment at a ride-sharing platform. *Management Science (forthcoming)*, https://doi.org/10.1287/mnsc.2022.4624.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753.

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

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review*, 80(3):313–336.

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. (2022). Just words? Just speeches? On the economic value of charismatic leadership. *Management Science*, 68(9):6355–6381.

Ärlebäck, J. B. and Albarracín, L. (2019). The use and potential of fermi problems in the stem disciplines to support the development of twenty-first century competencies. *ZDM Mathematics Education*, 51(6):979–990.

Ashraf, A. (2019). Do performance ranks increase productivity? Evidence from a field experiment. *CRC TRR190, Discussion Paper No. 196*.

Ashraf, N., Bandiera, O., and Lee, S. S. (2014). Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior & Organization*, 100:44–63.

Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, volume 1, pages 73–140. Elsevier.

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

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

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

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

Ball, S., Eckel, C., Grossman, P. J., and Zame, W. (2001). Status in markets. *The Quarterly Journal of Economics*, 116(1):161–188.

Bandiera, O., Barankay, I., and Rasul, I. (2007). Incentives for managers and inequality among workers: Evidence from a firm-level experiment. *The Quarterly Journal of Economics*, 122(2):729–773.

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

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

Bandiera, O., Prat, A., Hansen, S., and Sadun, R. (2020). CEO behavior and firm performance. *Journal of Political Economy*, 128(4):1325–1369.

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

Bardsley, N. (2000). Control without deception: Individual behaviour in free-riding experiments revisited. *Experimental Economics*, 3(3):215–240.

Barron, K., Ditlmann, R., Gehrig, S., and Schweighofer-Kodritsch, S. (2022). Explicit and implicit belief-based gender discrimination: A hiring experiment. *CESifo Working Paper 9731*.

Bartel, A. P., Beaulieu, N. D., Phibbs, C. S., and Stone, P. W. (2014). Human capital and productivity in a team environment: Evidence from the healthcare sector. *American Economic Journal: Applied Economics*, 6(2):231–59.

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: Industrial, military, and educational 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.

Baumeister, R. F. and Leary, M. R. (2017). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. In Canter, D., Laursen, B., and Zukauskiene, R., editors, *Interpersonal Development*, pages 57–89. Routledge, London.

Bellemare, C., Lepage, P., and Shearer, B. (2010). Peer pressure, incentives, and gender: An experimental analysis of motivation in the workplace. *Labour Economics*, 17(1):276–283.

Bennedsen, M., Pérez-González, F., and Wolfenzon, D. (2020). Do CEOs matter? Evidence from hospitalization events. *Journal of Finance*, 75(4):1877–1911.

Bertrand, M. and Schoar, A. (2003). Managing with style: The effect of managers on firm policies. *The Quarterly Journal of Economics*, 118(4):1169–1208.

Bigoni, M., Fort, M., Nardotto, M., and Reggiani, T. G. (2015). Cooperation or competition? A field experiment on non-monetary learning incentives. *The BE Journal of Economic Analysis & Policy*, 15(4):1753–1792.

Blader, S., Gartenberg, C., and Prat, A. (2020). The contingent effect of management practices. *Review of Economic Studies*, 87(2):721–749.

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

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.

Blimpo, M. P. (2014). Team incentives for education in developing countries: A randomized field experiment in Benin. *American Economic Journal: Applied Economics*, 6(4):90–109.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2013). Does management matter? Evidence from india. *The Quarterly Journal of Economics*, 128(1):1–51.

Bloom, N. and Van Reenen, J. (2007). Measuring and explaining management practices across firms and countries. *The Quarterly Journal of Economics*, 122(4):1351–1408.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2019a). Inaccurate statistical discrimination: An identification problem. *NBER Working Paper 25935*.

Bohren, J. A., Imas, A., and Rosenberg, M. (2019b). The dynamics of discrimination: Theory and evidence. *American Economic Review*, 109(10):3395–3436.

Bolton, P., Brunnermeier, M. K., and Veldkamp, L. (2013). Leadership, coordination, and corporate culture. *Review of Economic Studies*, 80(2):512–537.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, 109(3):739–73.

Borrego, C., Fernández, C., Blanes, I., and Robles, S. (2017). Room escape at class: Escape games activities to facilitate the motivation and learning in computer science. *Journal of Technology and Science Education*, 7(2):162–171.

Boss, V., Dahlander, L., Ihl, C., and Jayaraman, R. (2021). Organizing entrepreneurial teams: A field experiment on autonomy over choosing teams and ideas. *Organization Science (forthcoming)*, `https://doi.org/10.1287/orsc.2021.1520`.

Boudreau, L., Macchiavello, R., Minni, V., and Tanaka, M. (2021). Union leaders: Experimental evidence from Myanmar. *Working Paper*.

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

Bradler, C., Neckermann, S., and Warnke, A. J. (2019). Incentivizing creativity: A large-scale experiment with performance bonuses and gifts. *Journal of Labor Economics*, 37(3):793–851.

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.

Breza, E., Kaur, S., and Shamdasani, Y. (2018). The morale effects of pay inequality. *The Quarterly Journal of Economics*, 133(2):611–663.

Brown, J. (2011). Quitters never win: The (adverse) incentive effects of competing with superstars. *Journal of Political Economy*, 119(5):982–1013.

Brown, J. and Minor, D. B. (2014). Selecting the best? Spillover and shadows in elimination tournaments. *Management Science*, 60(12):3087–3102.

Bruhn, M., Karlan, D., and Schoar, A. (2018). The impact of consulting services on small and medium enterprises: Evidence from a randomized trial in mexico. *Journal of Political Economy*, 126(2):635–687.

Brunt, L., Lerner, J., and Nicholas, T. (2012). Inducement prizes and innovation. *Journal of Industrial Economics*, 60(4):657–696.

Burgess, S., Propper, C., Ratto, M., Kessler Scholder, S. v. H., and Tominey, E. (2010). Smarter task assignment or greater effort: The impact of incentives on team performance. *The Economic Journal*, 120(547):968–989.

Burks, S. V., Carpenter, J. P., Goette, L., and Rustichini, A. (2013). Overconfidence and social signalling. *Review of Economic Studies*, 80(3):949–983.

Bursztyn, L., González, A. L., and Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in Saudi Arabia. *American Economic Review*, 110(10):2997–3029.

Bursztyn, L. and Jensen, R. (2015). How does peer pressure affect educational investments? *The Quarterly Journal of Economics*, 130(3):1329–1367.

Bursztyn, L. and Jensen, R. (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, 9:131–153.

Card, D. and Krueger, A. (1994). Minimum wages and employment: A case study of the new jersey and pennsylvania fast food industries. *American Economic Review*, 84(4):772–793.

Card, D., Mas, A., Moretti, E., and Saez, E. (2012). Inequality at work: The effect of peer salaries on job satisfaction. *American Economic Review*, 102(6):2981–3003.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, 134(3):1163–1224.

Carlsson, F. and Martinsson, P. (2003). Design techniques for stated preference methods in health economics. *Health Economics*, 12(4):281–294.

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

Casas-Arce, P. and Martinez-Jerez, F. A. (2009). Relative performance compensation, contests, and dynamic incentives. *Management Science*, 55(8):1306–1320.

Casner-Lotto, J. and Barrington, L. (2006). Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century US workforce. *Available at ERIC (ED519465)*.

Castro, S., Englmaier, F., and Guadalupe, M. (2022). Fostering psychological safety in teams: Evidence from an RCT. *Available at SSRN 4141538*.

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

Charness, G. and Grieco, D. (2023). Creativity and corporate culture. *The Economic Journal*, 133(653):1846–1870.

Charness, G. and Kuhn, P. (2007). Does pay inequality affect worker effort? Experimental evidence. *Journal of Labor Economics*, 25(4):693–723.

Charness, G., Masclet, D., and Villeval, M. C. (2014). The dark side of competition for status. *Management Science*, 60(1):38–55.

Chen, R. and Chen, Y. (2011). The potential of social identity for equilibrium selection. *American Economic Review*, 101(6):2562–89.

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–57.

Chen, Y., Li, S. X., Liu, T. X., and Shih, M. (2014). Which hat to wear? Impact of natural identities on coordination and cooperation. *Games and Economic Behavior*, 84:58–86.

Chowdhury, S. M. (2021). The economics of identity and conflict. In *Oxford Research Encyclopedia of Economics and Finance*.

Chowdhury, S. M., Jeon, J. Y., and Ramalingam, A. (2016). Identity and group conflict. *European Economic Review*, 90:107–121.

Coffman, K. B., Exley, C. L., and Niederle, M. (2021). The role of beliefs in driving gender discrimination. *Management Science*, 67(6):3551–3569.

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.

Cullen, Z. and Perez-Truglia, R. (2022). How much does your boss make? The effects of salary comparisons. *Journal of Political Economy*, 130(3):766–822.

Dalenberg, S., Vogelaar, A. L., and Beersma, B. (2009). The effect of a team strategy discussion on military team performance. *Military Psychology*, 21(sup2):S31–S46.

De Paola, M., Gioia, F., and Scoppa, V. (2018). Teamwork, leadership and gender. *IZA Discussion Paper 11861*.

De Paola, M., Scoppa, V., and Nisticò, R. (2012). Monetary incentives and student achievement in a depressed labor market: Results from a randomized experiment. *Journal of Human Capital*, 6(1):56–85.

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

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

Delfgaauw, J., Dur, R., Sol, J., and Verbeke, W. (2013). Tournament incentives in the field: Gender differences in the workplace. *Journal of Labor Economics*, 31(2):305–326.

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

Deming, D. and Kahn, L. B. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, 36(1):337–369.

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

Dessein, W. (2007). Why a group needs a leader: Decision–making and debate in committees. *CEPR Discussion Paper 6168*.

Driskell, J. E., Salas, E., and Driskell, T. (2018). Foundations of teamwork and collaboration. *American Psychologist*, 73(4):334.

Drouvelis, M. and Nosenzo, D. (2013). Group identity and leading-by-example. *Journal of Economic Psychology*, 39:414–425.

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.

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

Dur, R., Kvaloy, O., and Schöttner, A. (2022). Labor-market conditions and leadership styles. *Managment Science*, 68(4):3150–3168.

Dustan, A., Koutout, K., and Leo, G. (2022). Second-order beliefs and gender. *Journal of Economic Behavior & Organization*, 200:752–781.

Dutcher, G. and Rodet, C. S. (2022). Which two heads are better than one? Uncovering the positive effects of diversity in creative teams. *Journal of Economics & Management Strategy*, 31(4):884–897.

Eckartz, K., Kirchkamp, O., and Schunk, D. (2012). How do incentives affect creativity? *CESifo Working Paper 4049*.

Eckel, C. C. and Grossman, P. J. (2005). Managing diversity by creating team identity. *Journal of Economic Behavior & Organization*, 58(3):371–392.

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

Englmaier, F., Grimm, S., Grothe, D., Schindler, D., and Schudy, S. (2021). The value of leadership: Evidence from a large-scale field experiment. *CESifo Working Paper 9273*.

Englmaier, F., Grimm, S., Grothe, D., Schindler, D., and Schudy, S. (2023a). The efficacy of tournaments for non-routine team tasks. *Journal of Labor Economics (forthcoming)*, `https://doi.org/10.1086/725553`.

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., Schindler, D., and Schudy, S. (2023b). Microaspects of leadership in non-routine analytical team tasks. mimeo.

Englmaier, F. and Leider, S. (2012). Contractual and organizational structure with reciprocal agents. *American Economic Journal: Microeconomics*, 4(2):146–83.

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

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

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

Erkal, N., Gangadharan, L., and Koh, B. H. (2021). Gender biases in performance evaluation: The role of beliefs versus outcomes. *Available at SSRN 3979701*.

Falk, A. and Kosfeld, M. (2006). The hidden costs of control. *American Economic Review*, 96(5):1611–1630.

Fershtman, C., Hvide, H. K., and Weiss, Y. (2006). Cultural diversity, status concerns and the organization of work. *Research in Labor Economics*, 24:361–396.

Fest, S., Kvaloy, O., Nieken, P., and Schöttner, A. (2019). Motivation and incentives in an online labor market. *CESifo Working Paper 7526*.

Fischbacher, U., Kübler, D., and Stüber, R. (2022). Betting on diversity – occupational segregation and gender stereotypes. *mimeo*.

Flynn, D., Nyhan, B., and Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38:127–150.

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

Fryer, R., Levitt, S., List, J., and Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. *NBER Working Paper 18237*.

Gächter, S. and Thöni, C. (2010). Social comparison and performance: Experimental evidence on the fair wage–effort hypothesis. *Journal of Economic Behavior & Organization*, 76(3):531–543.

Gerhart, B. and Fang, M. (2015). Pay, intrinsic motivation, extrinsic motivation, performance, and creativity in the workplace: Revisiting long-held beliefs. *Annual Review of Organizational Psychology and Organizational Behavior*, 2(1):489–521.

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

Gneezy, U. and List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, 118(3):1049–1074.

Gneezy, U. and Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, 94(2):377–381.

Gómez-Zará, D., Guo, M., DeChurch, L. A., and Contractor, N. (2020). The impact of displaying diversity information on the formation of self-assembling teams. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

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

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125.

Gross, D. P. (2020). Creativity under fire: The effects of competition on creative production. *Review of Economics and Statistics*, 102(3):583–599.

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 Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*, 16(2):250–279.

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

Hampole, M., Truffa, F., and Wong, A. (2021). Peer effects and the gender gap in corporate leadership: Evidence from MBA students. *Working Paper*.

Harrison, D. A. and Klein, K. J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review*, 32(4):1199–1228.

Harrison, D. A., Price, K. H., and Bell, M. P. (1998). Beyond relational demography: Time and the effects of surface-and deep-level diversity on work group cohesion. *Academy of Management Journal*, 41(1):96–107.

Harrison, D. A., Price, K. H., Gavin, J. H., and Florey, A. T. (2002). Time, teams, and task performance: Changing effects of surface-and deep-level diversity on group functioning. *Academy of Management Journal*, 45(5):1029–1045.

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

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

Hermalin, B. E. (1998). Toward an economic theory of leadership: Leading by example. *American Economic Review*, 88(5):1188–1206.

Heursen, L., Ranehill, E., and Weber, R. (2020). Are women less effective leaders than men? Evidence from experiments using coordination games. *CESifo Working Paper 8713*.

Hoffman, M. and Tadelis, S. (2021). People management skills, employee attrition, and manager rewards: An empirical analysis. *Journal of Political Economy*, 129(1):243–285.

Hole, A. (2017). Dcreate: Stata module to create efficient designs for discrete choice experiments, `https://EconPapers.repec.org/RePEc:boc:bocode:s458059`.

Hoogendoorn, S., Oosterbeek, H., and Van Praag, M. (2013). The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Management Science*, 59(7):1514–1528.

Hoogendoorn, S., Parker, S. C., and Van Praag, M. (2017). Smart or diverse start-up teams? evidence from a field experiment. *Organization Science*, 28(6):1010–1028.

Horwitz, S. K. and Horwitz, I. B. (2007). The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of Management*, 33(6):987–1015.

House, R. (1976). A 1976 theory of charismatic leadership. *Available at ERIC (ED133827)*.

House, R. J., Hanges, P. J., Ruiz-Quintanilla, S. A., Dorfman, P. W., Javidan, M., Dickson, M., and Gupta, V. (1999). Cultural influences on leadership and organizations: Project GLOBE. *Advances in Global Leadership*, Volume 1:171–233.

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.

Huber, L. R., Sloof, R., Van Praag, M., and Parker, S. C. (2020). Diverse cognitive skills and team performance: A field experiment based on an entrepreneurship education program. *Journal of Economic Behavior & Organization*, 177:569–588.

Hughes, A. M., Gregory, M. E., Joseph, D. L., Sonesh, S. C., Marlow, S. L., Lacerenza, C. N., Benishek, L. E., King, H. B., and Salas, E. (2016). Saving lives: A meta-analysis of team training in healthcare. *Journal of Applied Psychology*, 101(9):1266.

Imbens, G. W., Rubin, D. B., and Sacerdote, B. I. (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American Economic Review*, 91(4):778–794.

Isaksson, S. (2018). It takes two: Gender differences in group work. *Working Paper*.

Jäger, S., Roth, C., Roussille, N., and Schoefer, B. (2022). Worker beliefs about outside options. *NBER Working Paper 29623*.

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

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

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

Kennedy, J. A., Anderson, C., and Moore, D. A. (2013). When overconfidence is revealed to others: Testing the status-enhancement theory of overconfidence. *Organizational Behavior and Human Decision Processes*, 122(2):266–279.

Khan, B. Z. (2015). Inventing prizes: A historical perspective on innovation awards and technology policy. *Business History Review*, 89(4):631–660.

Kline, P. and Santos, A. (2012). A score based approach to wild bootstrap inference. *Journal of Econometric Methods*, 1(1):23–41.

Kluger, A. N. and DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2):254.

Kosfeld, M. and Neckermann, S. (2011). Getting more work for nothing? Symbolic awards and worker performance. *American Economic Journal: Microeconomics*, 3(3):86–99.

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

Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707.

Kozlowski, S. W. and Bell, B. S. (2013). Work groups and teams in organizations. Review update. *Handbook of Psychology*, 12:412–469.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

Kube, S., Maréchal, M. A., and Puppe, C. (2012). The currency of reciprocity: Gift exchange in the workplace. *American Economic Review*, 102(4):1644–62.

Kübler, D., Schmid, J., and Stüber, R. (2018). Gender discrimination in hiring across occupations: A nationally-representative vignette study. *Labour Economics*, 55:215–229.

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*.

Lazear, E. P. and Oyer, P. (2012). Chapter 12: Personnel economics. In *The Handbook of Organizational Economics*, pages 479–519. Princeton University Press.

Lazear, E. P. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):841–864.

Lazear, E. P., Shaw, K. L., and Stanton, C. T. (2015). The value of bosses. *Journal of Labor Economics*, 33(4):823–861.

Lindegaard, S. (2010). *The open innovation revolution: Essentials, roadblocks, and leadership skills*. John Wiley & Sons.

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

Littlepage, G. E., Hein, M. B., Moffett III, R. G., Craig, P. A., and Georgiou, A. M. (2016). Team training for dynamic cross-functional teams in aviation: Behavioral, cognitive, and performance outcomes. *Human Factors*, 58(8):1275–1288.

Lyons, E. (2017). Team production in international labor markets: Experimental evidence from the field. *American Economic Journal: Applied Economics*, 9(3):70–104.

Manski, C. F. (2001). Daniel McFadden and the econometric analysis of discrete choice. *The Scandinavian Journal of Economics*, 103(2):217–229.

Manthei, K., Sliwka, D., and Vogelsang, T. (2022). Talking about performance or paying for it? A field experiment on performance reviews and incentives. *Management Science (forthcoming)*, https://doi.org/10.1287/mnsc.2022.4431.

Marks, M. A., Mathieu, J. E., and Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, 26(3):356–376.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, 50(4):370.

Maximiano, S., Sloof, R., and Sonnemans, J. (2007). Gift exchange in a multi-worker firm. *The Economic Journal*, 117(522):1025–1050.

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

McEwan, D. and Beauchamp, M. R. (2014). Teamwork in sport: A theoretical and integrative review. *International Review of Sport and Exercise Psychology*, 7(1):229–250.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. New York, NY: Academic Press, Maastricht.

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

Meslec, N., Curseu, P. L., Fodor, O. C., and Kenda, R. (2020). Effects of charismatic leadership and rewards on individual performance. *Leadership Quarterly*, 31(6):101423.

Moldovanu, B., Sela, A., and Shi, X. (2007). Contests for status. *Journal of Political Economy*, 115(2):338–363.

Montagno, R. V. (1985). The effects of comparison others and prior experience on responses to task design. *Academy of Management Journal*, 28(2):491–498.

Morgan, J., Neckermann, S., and Sisak, D. (2020). Peer evaluation and team performance: An experiment on complex problem solving. Working Paper.

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.

Morikawa, T., Ben-Akiva, M., and McFadden, D. (2002). Discrete choice models incorporating revealed preferences and psychometric data. In *Advances in Econometrics*, pages 29–55. Emerald Group Publishing Limited.

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

NACE (2022). Job outlook: National association of colleges and employers.

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. *Leadership Quarterly*, 25(5):923–942.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

Niederle, M. and Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, 3(1):601–630.

Oldham, G. R., Nottenburg, G., Kassner, M. W., Ferris, G., Fedor, D., and Masters, M. (1982). The selection and consequences of job comparisons. *Organizational Behavior and Human Performance*, 29(1):84–111.

Patchen, M. (1958). The effect of reference group standards on job satisfactions. *Human Relations*, 11(4):303–314.

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. *Leadership Quarterly*, 13(5):561–581.

Ramm, J., Tjotta, S., and Torsvik, G. (2013). Incentives and creativity in groups. *CESifo Working Paper 4374*.

Restivo, M. and Van De Rijt, A. (2012). Experimental study of informal rewards in peer production. *PloS one*, 7(3):e34358.

Rigdon, J. and Hudgens, M. G. (2015). Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine*, 34(6):924–935.

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.

Salas, E., Tannenbaum, S. I., Kozlowski, S. W., Miller, C. A., Mathieu, J. E., and Vessey, W. B. (2015). Teams in space exploration: A new frontier for the science of team effectiveness. *Current Directions in Psychological Science*, 24(3):200–207.

Sarsons, H. (2017). Recognition for group work: Gender differences in academia. *American Economic Review*, 107(5):141–45.

Sarsons, H., Gërxhani, K., Reuben, E., and Schram, A. (2021). Gender differences in recognition for group work. *Journal of Political Economy*, 129(1):101–147.

Schram, A., Brandts, J., and Gërxhani, K. (2019). Social-status ranking: A hidden channel to gender inequality under competition. *Experimental Economics*, 22(2):396–418.

Scotchmer, S. (2004). *Innovation and Incentives*. MIT Press.

Sen, A. (2007). *Identity and violence: The illusion of destiny*. Penguin Books India.

Spencer, S. J., Steele, C. M., and Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 35(1):4–28.

Stewart, G. L. (2006). A meta-analytic review of relationships between team design features and team performance. *Journal of Management*, 32(1):29–55.

Tajfel, H. and Turner, J. (2001). An integrative theory of intergroup conflict. In Hogg, M. A. and Abrams, D., editors, *Key Readings in Social Psychology*, pages 94–109. Psychology Press, New York.

Tan, J. H. and Bolle, F. (2007). Team competition and the public goods game. *Economics Letters*, 96(1):133–139.

Terwiesch, C. and Ulrich, K. T. (2009). *Innovation tournaments: Creating and selecting exceptional opportunities*. Harvard Business Press.

Terwiesch, C. and Xu, Y. (2008). Innovation contests, open innovation, and multiagent problem solving. *Management Science*, 54(9):1529–1543.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309.

Toma, M. and Bell, E. (2021). Understanding and improving policymakers' sensitivity to program impact. *Available at SSRN 4435532*.

Torrance, E. P. (1966). *Torrance tests of creative thinking: Norms-technical manual*. Princeton: Personnel Press.

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

Van Dick, R., Grojean, M. W., Christ, O., and Wieseke, J. (2006). Identity and the extra mile: Relationships between organizational identification and organizational citizenship behaviour. *British Journal of Management*, 17(4):283–301.

Van Knippenberg, D. (2000). Work motivation and performance: A social identity perspective. *Applied Psychology*, 49(3):357–371.

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.

Weidmann, B. and Deming, D. J. (2021). Team players: How social skills improve team performance. *Econometrica*, 89(6):2637–2657.

Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039.

Ye, T., Ai, W., Chen, Y., Mei, Q., Ye, J., and Zhang, L. (2022). Virtual teams in a gig economy. *Proceedings of the National Academy of Sciences*, 119(51):1–12.

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