I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17165

How Do You Find a Good Manager?

Ben Weidmann
Joseph Vecci
Farah Said
David Deming
Sonia Bhalotra

JULY 2024

I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17165

How Do You Find a Good Manager?

Ben Weidmann
Harvard Kennedy School

Joseph Vecci
University of Gothenburg and IZA

Farah Said
Lahore University of Management Sciences

David Deming
Harvard Kennedy School and NBER

Sonia Bhalotra
University of Warwick and IZA

JULY 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

How Do You Find a Good Manager?*

This paper develops a novel method to identify the causal contribution of managers to team performance. The method requires repeated random assignment of managers to multiple teams and controls for individuals' skills. A good manager is someone who consistently causes their team to produce more than the sum of their parts. Good managers have roughly twice the impact on team performance as good workers. People who nominate themselves to be in charge perform worse than managers appointed by lottery, in part because self-promoted managers are overconfident, especially about their social skills. Managerial performance is positively predicted by economic decision-making skill and fluid intelligence – but not gender, age, or ethnicity. Selecting managers on skills rather than demographics or preferences for leadership could substantially increase organizational productivity.

JEL Classification: M54, J24, C90

Keywords: teamwork, managers, skills, measurement, experiment

Corresponding author:
Ben Weidmann
Harvard Kennedy School
79 John F. Kennedy St
Cambridge, MA 02138
USA
E-mail: benweidmann@hks.harvard.edu

* Funding for this research was provided by Bhalotra as Co-Investigator of the Centre for Microsocial Change (ESRC Award ES/S012486/1) at the University of Essex, supplemented by funds awarded to Deming and Weidmann by Schmidt Futures and to Vecci by the University of Gothenburg. We thank the Essex Lab research assistance and lab manager for their help in implementing the experiment. Ethics approval was provided by the University of Warwick (HSSREC 163/21-22). The Pre-Analysis Plan was submitted to the AEA registry prior to data collection (AEARCTR-0012258). Weidmann led the experimental design and data analysis, and contributed to writing and data collection. Vecci contributed to the experimental design, data collection, analysis, writing and contributed financially. Said contributed to the experimental design, analysis and data collection. Deming and Bhalotra contributed to the writing, experimental design and financially.

1 Introduction

Management matters greatly for economic performance (Bloom et al., 2013, 2016). There are large and persistent productivity differences between managers within firms and across countries (Lazear et al., 2015; Hjort and Chandler, 2022). Good managers increase productivity through many channels, including motivation, monitoring performance, and reallocating workers to better roles or job tasks (Metcalfe et al., 2023; Adhvaryu et al., 2023; Minni, 2023; Fenizia, 2022).

How can firms identify good managers? In practice, firms rely heavily on the judgment of existing managers, which suffers from well-known biases (Kahneman and Klein, 2009; Hoffman et al., 2017; Chamorro-Premuzic, 2019; Feld et al., 2022). Firms can also select managers based on personality traits, education and cognitive ability, which have been shown to positively predict performance (Kaplan, 2012; Hoffman and Tadelis, 2021; Queiro, 2022).

However, existing work on manager selection suffers from two issues. First, managers are not randomly assigned to teams, which makes it difficult to causally identify managerial performance. Second, managers in the field are a highly non-random sample, which may lead to incorrect inferences about the characteristics that truly improve performance. For example, Benson et al. (2019) find that managers are selected based on their performance as line workers (the "Peter Principle", where employees are promoted to their level of incompetence) even though other attributes are better predictors of managerial performance. Selection based on the Peter Principle induces a spurious negative correlation between worker performance and managerial performance, even though the marginal effect of worker performance is likely positive.

In this paper we introduce a new experimental method to identify the causal contribution of managers to team performance. The method requires repeated random assignment of managers to multiple teams and careful controls for individual performance predictors (Weidmann and Deming, 2021). Intuitively, good managers consistently cause their workers to exceed predicted performance.

We estimate managerial performance in a large, pre-registered lab experiment using a novel Collaborative Production Task that emulates real-world team production by requiring managers to coordinate, monitor and motivate workers. After controlling for individual performance, a manager whose estimated performance is one standard deviation above the average improves team performance by about 0.23 standard deviations. This highlights the critical role of effective management in enhancing team productivity, showing that a good manager is almost twice as valuable as a good worker.[1]

[1]We find that a one standard deviation increase in worker production skills increases team performance

Do people who want to be managers perform well in the job? We explore this question by randomly varying the manager selection mechanism in our experiment. After describing the expected tasks and compensation structure of the manager and worker roles, we elicit participants' eagerness to be a manager on a 1-10 scale. Half of groups were randomly assigned to a "self-promotion" treatment where participants with the strongest preferences became managers. Managers were assigned randomly in the other half of groups. We find that self-promotion is worse than choosing managers randomly. Teams with self-promoted managers perform 0.1 standard deviations lower than teams with randomly assigned managers. This magnitude is roughly equivalent to being assigned a manager with fluid IQ one standard deviation lower.

We show that self-selection can lead to mistaken inferences about the characteristics of good managers. People who prefer to be in charge – who we call 'self-promoters' – have characteristics that differ from the broader population. For example, we find suggestive evidence that self-promoters tend to overestimate their own social skills relative to an objective test of emotional perceptiveness called the Reading the Mind in the Eyes Test (RMET).[2] Among self-promoted managers, we find a negative relationship between self-reported people skills and managerial performance. In contrast, randomly selected managers do not tend to overestimate their social skills, and we find no negative relationship between self-reported people skills and managerial performance.

We simulate the impact of different managerial selection rules using results from the random assignment arm of the experiment. We find that selecting managers based on economic decision-making skill improves performance by 0.7 standard deviations relative to self-promotion. This is equivalent to replacing an average worker with a worker in the 99th percentile of individual productivity. Selecting managers based on economic decision-making skill and fluid intelligence yields significantly better performance than the random assignment benchmark, selection on social skills, or selection on worker task skill (e.g. the "Peter principle").

What are good managers doing to help their team succeed? We find that good managers monitor their workers to avoid wasted effort, match workers to the right tasks, and keep workers motivated and engaged. We show that some groups finish the task with workers

by 0.26 standard deviations, compared to the 0.23 standard deviation impact of a good manager. Since our experiment is conducted in three-person teams with one manager and two workers, the results imply that a good manager is almost twice as valuable as a good worker. Consistent with this interpretation, Weidmann and Deming (2021) find that good team players improve group performance in three-person teams by 0.13 standard deviations, a bit more than half of the manager effect estimated here. Similarly, in their field study Lazear et al. (2015) estimate that a good boss is 75% more valuable than a good employee.

[2]We also find that self-promoters are generally more overconfident in their performance and abilities. This is consistent with related evidence that managers and executives tend to be overconfident (Malmendier and Tate, 2015).

assigned to sub-tasks that, even if they were completed, would not improve the group's total score - suggesting wasted effort. Managers who are 1 SD above average in terms of estimated performance waste worker effort only half as much as other managers (8 percent vs. 16 percent overall). We also find that good managers are much more likely to optimize task assignments according to workers' comparative advantage. Finally, we find that manager performance is strongly influenced by how teams perform at the end of the task: teams led by a manager who is 1 SD above average solve 0.6 more problems in the final two minutes, whereas the performance gap in the first two minutes is only 0.3 problems.

Our paper makes four main contributions. First, we develop a new method for estimating a manager's ability to improve the output of the teams they supervise. The method requires both repeated random assignment to teams, and controls for individual skill. Importantly, this approach offers the possibility to *prospectively* identify good managers. Our approach can be easily implemented in the field (e.g., Falk and Heckman 2009; Charness and Kuhn 2011). There are several reasons to believe that our results may hold in more realistic settings. Herbst and Mas (2015) find that experiments on peer productivity spillovers yield very similar magnitudes when conducted in the lab versus the field. Also, our estimates of the productivity value of a good manager are very similar to Lazear et al. (2015), who estimate the impact of managers on individual worker performance using data from a large employer.[3]

Second, we quantify the impact of manager selection policies. In the US, the most common approach to manager selection is a combination of self-promotion and the judgment of existing managers - who themselves may not have been optimally selected (Chamorro-Premuzic, 2019). By randomly assigning the method of manager selection *in addition* to identifying the causal effect each manager has on their team, our design enables us to quantify the large potential benefits of moving towards a skills-based hiring approach. Adopting objective, skills-based selection methods can significantly enhance managerial effectiveness and overall team performance.

Third, we illuminate the ways in which good managers matter, in particular the importance of task allocation and comparative advantage. The single best predictor of managerial performance is a theoretically grounded measure of allocative efficiency that

[3]There are two important differences between our study and Lazear et al. (2015). First, they identify manager effects through nonrandom worker rotation rather than random assignment. Second, in their setting manager-worker relationships are dyadic, so there is no role for task allocation or interaction between workers. This makes it difficult to distinguish spurious factors from the genuine characteristics associated with manager success. Our study also differs from lab experiments examining the impact of leaders in public goods games (Cooper et al., 2020a; Brandts and Cooper, 2007; Brandts et al., 2015; Bhalotra et al., 2022; Potters et al., 2007; Güth et al., 2007a). These studies typically limit the leader's role to simple choices like making the first move in a public goods game, whereas our study estimates manager effects in more complex tasks.

Caplin et al. (2024) call *economic decision-making skill.* This suggests that systematic evaluation of managers' skills would increase organizational performance relative to selection on demographic characteristics or on a worker's belief about their suitability for management.

Fourth, we contribute to the broader literature on the determinants of effective leadership and the importance of how leaders are selected. Alchian and Demsetz (1972) argue that whenever production occurs in teams, leaders are necessary for monitoring performance and preventing free-riding. Leaders can also help solve coordination problems (Brandts and Cooper, 2007; Güth et al., 2007b; Sahin et al., 2015). Consistent with these findings, Englmaier et al. (2024) show that non-hierarchical teams improve their performance when they are randomly encouraged to select a leader. Many studies find that leaders have a large impact on team performance, and that how leaders are selected affects workers perceptions and their productivity (e.g. Bennedsen et al., 2007; Bloom and Van Reenen, 2007; Nevicka et al., 2011; Desseranno et al., 2019; Cooper et al., 2020b). Like several other studies, we find that women are much less likely to nominate themselves for leadership roles despite being equally or more effective (Reuben et al. 2010; Ertac and Gürdal 2012; Chakraborty and Serra 2023; Born et al. 2022). This literature struggles to distinguish between the influence of leaders from that of team members, making it challenging to understand both the determinants of effective leadership and the impact of different selection mechanisms on team performance. Our study addresses this gap by identifying and quantifying managers' causal contributions to team performance under various selection mechanisms.

The paper proceeds as follows. Section 2 describes the experiment and the data. Section 3 develops our identification and measurement strategy. Section 4 presents the main results. Section 5 explores mechanisms, and Section 6 concludes.

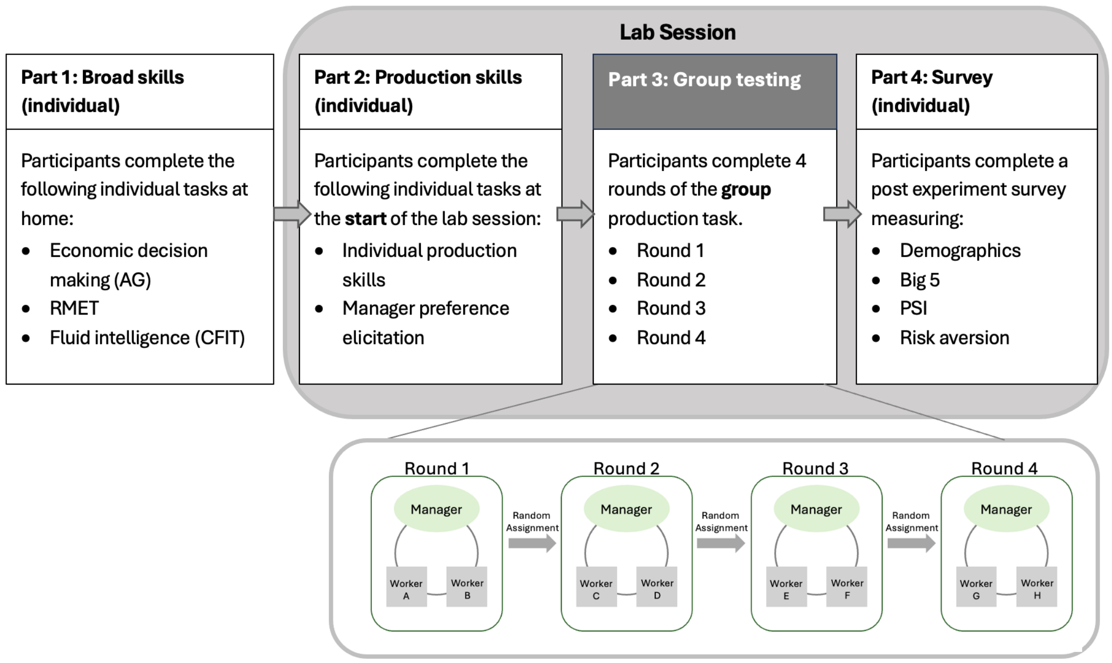
2 Description of experiment and data

2.1 Overview of the experiment

The experiment involved both individual and group testing. Individual testing focused on three areas. First, we measured potential predictors of manager skill including fluid intelligence (CFIT); emotional perceptiveness (RMET, Baron-Cohen et al. 2001); economic decision-making skills (Caplin et al., 2024); and participants' preference for being a manager. Second, we measured broad personality and demographic characteristics including the Big 5 personality inventory, ethnicity, age and levels of work experience and education. Third, we measured each participant's production skills. These were assessed with a set of tests designed to be individual analogues of the group task that participants subsequently worked on in the face-to-face lab sessions.

Group assessments took place at the Essex lab in the UK. Each group consisted of a manager and two workers.[4] At the start of group testing, participants were assigned the role of 'manager' or 'worker'. Roles were retained throughout the experiment. Each group of three people completed the Collaborative Production Task, a novel task in which the group's goal is to solve problems across three different modules: numerical, spatial and analytical reasoning. In the Collaborative Production Task, managers are responsible for assigning workers to different modules, monitoring group progress and keeping their team engaged. The Task is described in detail in Section 2.4. Throughout the experiment each participant worked in four randomly assigned groups. An overview of the sequence of the tasks can be found in Figure 1.

Figure 1: Experimental design flow chart



Notes: the figure describes the flow of the experiment from the perspective of participants. In the group testing phase each 'round' involves working in one group of three people. Each round involves a parallel version of the Collaborative Production Task, described in Section 2.4. The RMET is the Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001). Economic-decision making is measured by the Assignment Game (Caplin et al., 2024). CFIT is the Culture Fair Intelligence Test. PSI is the Political Skill Inventory, and Risk Aversion is a single question asking about risk preferences. For more information on measures see Section 2.

Lab sessions involved 12, 15 or 18 participants, and lasted around two hours in total. Each session was randomly assigned to one of two conditions that governed the way

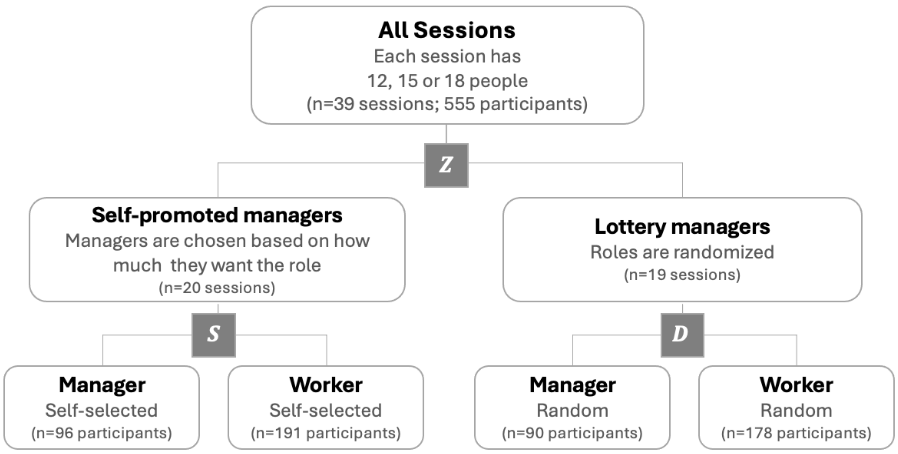
[4]In the experimental materials we referred to 'workers' as 'team members'. See the online Appendix B for a full description of experimental materials.

in which managers were selected: self-promotion and lottery. Before the group testing began all participants rated their preference for being a manager on a scale of 1-10. They were provided detailed information about the role of the manager (see Online Appendix for screenshots). Participants were informed that managers would be responsible for directing the group, communicating with team members, delegating, and motivation. Participants were also briefed on the incentive structure for managers, which is described in detail in Section 2.5. In the self-promotion condition, the role of manager was assigned to participants with the strongest preference for being in charge. In the lottery condition roles were randomly assigned.

2.2 Randomization and design

Our design has two levels of randomization, as summarized in Figure 2. First, we randomize the manager assignment mechanism. Then we repeatedly randomly assign participants to groups of three people. Lab sessions were randomly assigned to have either 'self-promoted managers' or 'lottery managers'. In total we had 39 sessions: 20 with self-promoted managers and 19 with managerial lotteries.

Figure 2: Randomization scheme



Notes: the figure describes the participant flow and randomization scheme. Z is a random variable that determines the way in which managers will be assigned. S is a self-selection mechanism, based on participants preferences for being a manager. D is a random variable that assigns 1/3rd of participants to be a manager in a lottery.

Within each session, participants were randomly assigned to groups so that each group had one manager and two workers. Over the course of the experiment each participant was assigned to four groups. To minimize the chances that the same people worked together multiple times, we followed a randomization scheme that minimized repeat interactions with the same team members. In sessions with 15 and 18 the scheme ensured that managers never worked with the same worker more than once.[5]

2.3 Individual tasks

2.3.1 Measuring individual productivity

Before group testing began, we assessed individuals' ability to solve problems on their own. The group task involved three types of problems (numerical, spatial, and analytical reasoning) so participants were asked to complete *individual* assessments in each of these domains beforehand.[6] This provided us with a set of measures we could use to predict the productive skill of any randomly assembled group.

The numerical reasoning test assessed the ability to understand and manipulate number sequences. Participants were asked to fill in a missing number based on a numerical pattern. An example item is presented in panel A of Figure 3. The spatial reasoning test evaluated the capacity to manipulate and conceptualize objects in two or three dimensions. For instance, participants were shown a simple three-dimensional image and asked how it might look if it were duplicated and rotated (see panel B of Figure 3). Last, the analytical reasoning test assessed the ability to analyze language problems and to understand analogies. This module relied heavily on analogical questions and vocabulary questions focused on antonyms or synonyms. (see example in panel C of Figure 3).

For each of the three tests, participants were given four minutes to solve as many problems as possible. They received 1 point for each correct answer and lost 0.5 points for each incorrect answer. Participants were aware of this scoring rule. We made this design choice to discourage guessing.

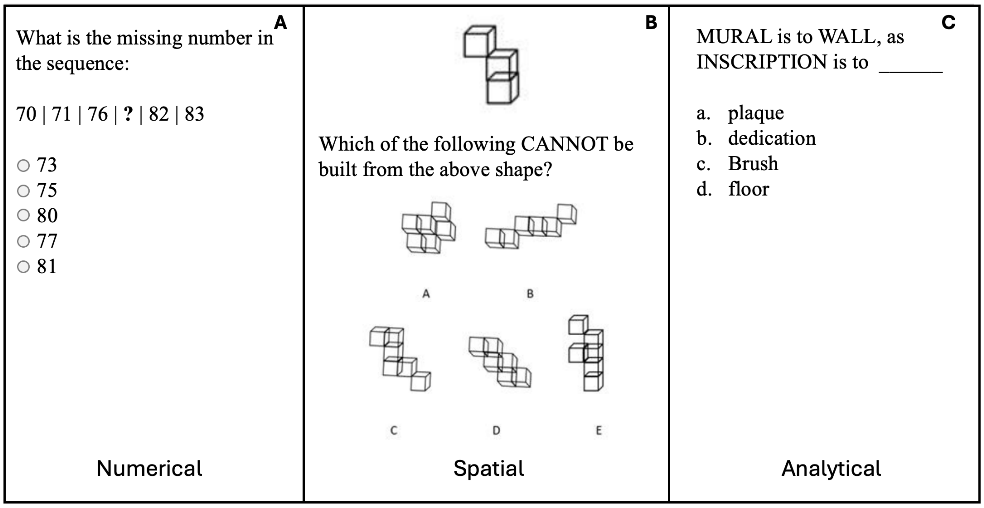
2.3.2 Broad measures of individual skill

Fluid intelligence was measured with a form of the Culture Fair Intelligence Test (CFIT III), a widely-used assessment of the ability to solve novel problems. An example item is provided in panel A of Figure 4. We measured social skill using the Reading the Mind in the Eyes (RMET, (see Baron-Cohen et al., 2001). This psychometrically validated

[5]In sessions with 12 participants, our randomization scheme ensured that people would work together a maximum of two times.

[6]We chose these three domains in part because previous research suggested relatively low cross-correlation among them (e.g. Chabris, 2007; Haier et al., 2009). In our study, we found similar correlations in individual scores across domains: estimated correlation coefficients were between the numerical, spatial and analytical scores are between 0.16 and 0.19 (n=555). This is a useful property in the context of the group task, as it makes allocation decisions more consequential. Items included in the tests form a crucial component of job assessments such as the Wonderlic battery of tests, Johnson O'Connor battery, Criteria Cognitive Aptitude Test, and the SHL Verify Ability Tests.

Figure 3: Example items measuring individual ‘productive skills’



Notes: panel A shows an example *numerical* item; panel B a *spatial* item; panel C an *analytical* item. During individual testing, participants were tested on each problem type sequentially (4 minutes each for numerical; spatial and analytical)..

test of emotional perceptiveness assesses emotional perceptiveness by presenting participants with photos of faces, cropped so that only the eyes are visible. For each set of eyes, participants are asked to choose which emotion, from four options, best describes the emotion being expressed. An example item is presented in panel B of Figure 4. Definitions of all the words were available via links to an online dictionary.[7]

We measured economic decision-making skill, defined by Caplin et al. (2024) as the ability to make good resource allocation decisions. This was assessed using the Assignment Game (Caplin et al., 2024). In the Assignment Game participants play the role of a manager who assigns fictional workers to tasks. To perform well, participants must deal with an attentionally-demanding numerical environment, understand comparative advantage intuitively, and avoid biases that undermine numerical decision making (e.g. anchoring). A screenshot from the game is presented in panel C of Figure 4.

2.3.3 Self-reported measures of personality and working styles

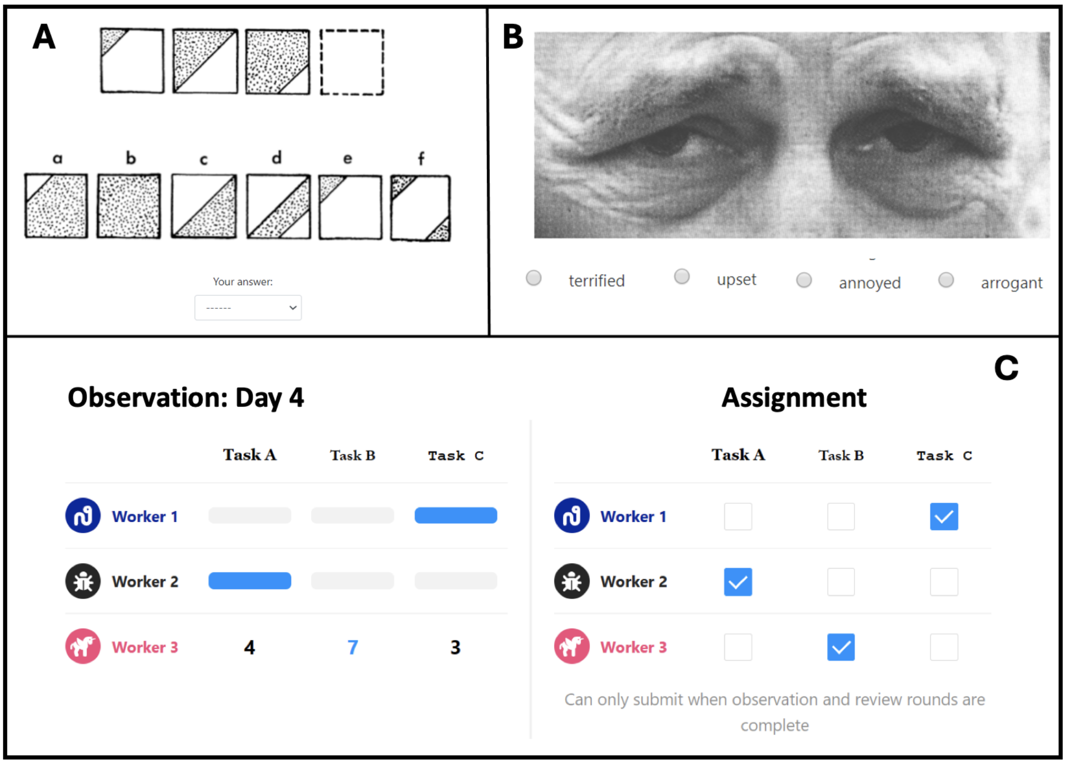
Participants completed the 10-item version of the Big 5 personality inventory (Gosling et al., 2003). Big 5 personality traits typically have positive correlations with job performance (e.g. Hurtz and Donovan, 2000) and with performance in laboratory studies of small-group problem-solving (Bell, 2007). Participants also completed the shortened

[7] We used a slightly shortened version of the test (26 items rather than 36, as per Weidmann and Deming (2021). We removed 5 female and 5 male images primarily to save time in what was becoming an onerous battery of tests.

Political Skill Inventory (PSI) described in Ferris et al. (2005).[8] The PSI measures "the ability to effectively understand others at work, and to use such knowledge to influence others to act in ways that enhance one's personal and/or organizational objectives" (Ahearn et al., 2004).

Finally, we measured risk appetite, based on the question: "[a]re you generally a person who is fully prepared to take risks, or do you try to avoid taking risks? Please choose a number on a scale from one (unwilling to take risks) to ten (fully prepared to take risks)". The internal and external validity of this measure has been extensively documented in previous studies (e.g. Dohmen et al., 2011; Becker et al., 2012; Hardeweg et al., 2013). For instance, Dohmen et al. (2011) show that the measure is strongly predictive of actual risky behaviour.

Figure 4: Example items from broad tests of individual skill



Notes: Panel A is an example item from CFIT, a measure of fluid intelligence. Panel B is an example item from RMET, a measure of emotional perceptiveness; the correct answer is 'upset'. Panel C is a screenshot from the Assignment Game. For a full description of the game, see Caplin et al. (2024).

[8]Following our pre-analysis plan, items that related to the "networking" subscale are removed, as these are not relevant to our lab setting.

2.4 Group task: Collaborative Production Task

The group task was designed to satisfy four criteria. First, the task had right and wrong answers to allow for objective scoring and to reduce measurement error. Second, collaboration was essential for group success in the task, a feature which is often lacking in group tasks (Larson, 2013). Third, we sought a group task that had an individual analogue, which would allow us to control for differences in each team's endowment of individual task-specific skill. Finally, given our focus on managers, the task had to have a clear and distinct role for managers which replicated real-world managerial demands such as coordination, delegation, and motivation. Based on these criteria we created the Collaborative Production Task.

In the task, groups are required to work on three question modules: numerical, spatial and analytical reasoning.[9] The group receives a score for each module based on how many problems they have solved. They receive one point for a correctly solved problem and lose 0.5 points for an incorrect solution. Each person in the team works on their own computer trying to solve a different problem. Crucially, the manager decides who will be working on each module. The overall team score is the *minimum* module score. This is similar to the weakest link coordination game, where collaboration is essential for success (e.g. Hirshleifer 1983).[10]

An alternative is to define the group score as the 'total number of problems solved correctly'. An issue with this rule is that if one group has a team member who is strong on a particular dimension then they can carry the team with no effort or input from others. Since individual skills are imprecisely measured, as in many real-world scenarios, the residual in our conditioning model (discussed below) will reflect unmeasured individual skill rather than collaboration. Additionally, with a 'total correct' scoring rule, once the best performer for each task is identified and a good allocation is made, the manager's role is much more narrowly focused on production, rather than communication and dynamic decision-making. Our chosen scoring rule increases the need for managers to monitor, communicate and make decisions on the fly, mirroring real-life scenarios in which managers need to respond to changing demands.

Each group of three sat next to each other in the lab, with the manager in the middle

[9]Although the questions were analogous to those in the individual tests of numeracy, spatial and analytical reasoning, there was no overlap in the specific problems we used.

[10]This setup represents a scenario where team production relies on the contributions from all components of production. For instance, in producing a report, the final product is only complete when all sections are combined. In manufacturing, a single malfunctioning component can halt the entire production line, while in project management, delays in one segment can affect the overall project timeline. A manager is generally required to coordinate across teams to ensure the successful production of the combined product. Therefore, the weakest-link set up is relevant for practical purposes because such coordination is common in various economic and social settings (Camerer and Weber, 2013; Riedl et al., 2015; Gächter et al., 2023).

and a worker on either side. There were no barriers and teammates could easily talk to each other throughout the experiment. In the instructions for this task, group members were informed multiple times that it was possible to talk to group members. To avoid cross team communication or interference, teams were separated from other teams by barriers and spare computer terminals.

Overall, each group worked together for around 15 minutes. At the start of the task, participants were given time to introduce themselves. In the first round of the experiment participants were presented with detailed instructions, including multiple comprehension checks. After introductions, groups could spend time strategizing about how they wanted to tackle the task. After every manager in the session had entered their initial task assignments (i.e. after they had decided who should work on different modules) the timer began and the first set of questions were shown. Groups worked on problems for 8 minutes in total. This included two 'break periods' of 1 minute, in which managers had time to regroup and motivate their team and/or re-strategize. Managers were encouraged to communicate with their team throughout the experiment and could reassign their teammates to different modules at any time after the task had started.

2.4.1 The role of the manager

Managers had several distinct responsibilities, including delegation, monitoring and motivation. First, *managers were responsible for deciding who did what.* Managers were allowed to delegate in any way they saw fit, provided everyone (including the managers themselves) were allocated to a module. Managers were able to allocate more than one person to any given module.[11] If, for instance, a manager allocated all three participants to the 'numerical' module, each participant would work on a different number problem. Allocations were fully dynamic and managers could change their allocation at any time. Before the timer began, managers were presented with full information about each of their team-members' individual scores on the individual assessments of numerical, spatial and analytical tests (conducted prior to the group session). Managers were able to review this information about the skill profile of their teammates at any point during the Collaborative Production Task.

Second, *managers monitored progress* throughout the task. This was especially important given the 'minimum effort' scoring rule. Only the manager's computer terminal showed the overall team score (i.e. the minimum module score). However, managers were *not* told which module has the lowest score. Consequently, managers needed to talk with their teammates to find out which modules needed attention. This required communication and strong situational awareness as managers could, and often did, get

[11]With three people and three modules, this allowed for 27 possible allocations.

swept up working on the module they assigned to themselves.

Third, *managers motivated their team* throughout the task. This is important given that the underlying tasks are both somewhat repetitive and cognitively demanding. As noted below (Section 2.5) the incentive structure meant that managers were unable to rely on financial incentives to motivate team members.

These responsibilities were specifically incorporated into the task as they are common managerial duties. For example, monitoring is a key management practice measured in the World Management Survey (Bloom et al., 2014). Furthermore, these responsibilities markedly differ from the existing lab-based literature on managerial/leader impact (Cooper et al., 2020a; Brandts and Cooper, 2007; Brandts et al., 2015; Bhalotra et al., 2022). In this literature, the leader or manager typically acts as the first mover in public goods or coordination games. In these settings, the manager does not engage in delegating, monitoring, or motivating team members.

2.5 Recruitment, sample and incentives

Participants were recruited from the Essex University Economics Lab sample pool. Column 1 of Table 1 reports descriptive statistics of the overall sample. Our sample comprises 555 individuals, forming a total of 728 groups of three across the four rounds.[12] 46% of the sample were female, with an average age of 25. The sample was ethnically diverse, with a majority of participants identifying as Asian or Asian British (54%). The median participant was a graduate student with two years of work experience. Columns 2 and 3 report sample statistics for the two treatment arms (self-promoted and lottery). Column 4 presents the results of balance tests across the two arms. None of the characteristics have mean differences that are significantly different from zero at the 5% level.

[12] We dont have data for 12 groups due to data errors, primarily stemming from one session where the internet cut out during the final round. For this session the 4th round data was not collected.

Table 1: Sample and balance

	Overall sample (1)	**Self-promoted arm (2)**	**Lottery arm (3)**	**p-value (4)**
Demographics				
Female (%)	46.3%	49.1%	43.3%	0.17
Age mean (yrs)	25.0	25.2	24.9	0.37
Work experience mean (yrs)	2.5	2.5	2.5	0.98
Asian or Asian British (%)	53.9%	52.6%	55.3%	0.54
White (%)	18.3%	17.8%	18.8%	0.76
Black, Caribbean or African (%)	15.5%	17.0%	13.9%	0.32
Other ethnic identity*	12.3%	12.6%	12.0%	0.10
Graduate students (%)	67.6%	70.0%	65.0%	0.22
Skill assessments				
Task skills	0.00	-0.01	0.01	0.86
Fluid intelligence (CFIT)	0.00	-0.01	0.01	0.81
Economic Decision-making (AG)	0.00	-0.04	0.05	0.30
Emotional perceptiveness (RMET)	0.00	-0.07	0.08	0.08
Personality and working styles				
Extraversion (Big5)	0.00	-0.01	0.01	0.90
Openness (Big5)	0.00	-0.05	0.05	0.25
Agreeableness (Big5)	0.00	-0.01	0.01	0.77
Neuroticism (Big5)**	0.00	0.05	-0.05	0.30
Conscientiousness (Big5)	0.00	-0.01	0.01	0.86
Political Skill Inventory (PSI)	0.00	-0.05	0.05	0.30
Indecisiveness Index (II)	0.00	0.01	-0.01	0.82
Risk appetite***	0.00	0.03	-0.03	0.49
Count	555	287	268	-

Notes: Skill assessments and personality measures are all standardized to have mean=0, sd=1. *Other ethnic identity includes 'Mixed or multiple ethnic groups', 'Other ethnic group' and people who preferred not to say. **Neuroticism (Big5) is reverse coded. P-values come from t-tests comparing means in the 'lottery' and 'self-selection' arms. *** Risk appetite refers to the willingness to take risks.

2.5.1 Incentives

Participants who completed the study were paid £35 on average, with a minimum payment of £29 and a maximum of £41. The individual tasks were incentivized with a bonus of £0-£4. Managers received a flat payment of £25 at the end of the session for completing all the experimental tasks. In the Collaborative Production Task, managers received a bonus of £4-£12 that depended on team performance in one randomly selected round. Managers in the top 40% of performers were paid a bonus £12 while those in the bottom 40% received £4. Other managers were paid £8. The average manager was paid £33 for the group session.

Workers did not have performance incentives for the group tasks and were paid a fixed rate of £33 for the group session. This was the same average payment as the manager. We chose to have different incentive structures for managers and workers for three reasons. First, managers in many organizations face steeper performance incentives than workers. Second, in many occupations, workers receive a fixed salary that does not significantly depend on their marginal effort. This is often because it is difficult to observe a worker's individual contribution to the overall team performance. Third, we wanted to allow for the possibility of managers motivating their team without having to rely on the motivation of financial incentives to perform.

2.6 Eliciting manager preferences

We elicited preferences for being a manager in both arms of the experiment. We began by describing the role of the manager in the upcoming group task as someone who would be responsible for delegating, coordinating and making decisions. We then emphasized that managers and workers would get paid the same on average. Finally, we encouraged participants to choose the role 'that best fits your skills'. We then asked participants 'how much do you want to be the manager?' on a scale of 1 to 10, where 1 is 'I really DON'T want to be manager' and 10 is 'I really DO want to be manager'. The average participant spent more than a minute deciding on their preference.[13].

Figure 5 presents the distribution of manager preferences among managers for both arms of the experiment. The left panel shows the distribution in the lottery arm. The distribution is fairly uniform, suggesting that neither role was dominantly desirable (mean response = 6, sd = 3). The left panel shows the distribution of preferences among managers in the self-promotion arm. By design, managers in this arm strongly

[13]We tested whether participants spent more time on this decision in the 'self-promotion' arm of the experiment, but found that participants in the lottery arm spent slightly more time on average (mean difference = 11 seconds, p=0.04). Participants did not know which treatment they were in when making this decision

prefer to be in charge, with almost half the managers responding with a 10 on the 1-10 scale.

Figure 5: Histogram of preference to be manager, by treatment arm

Notes: Plot represents counts (n=96 self-promoted managers; n=90 lottery managers). These are participants' responses to the question, 'how much do you want to be manager' on a scale of 1-10, where 1 is 'I really DON'T want to be manager' and 10 is 'I really DO want to be manager'.

2.7 Who wants to be a manager?

Table 2 explores the associations between various individual characteristics and wanting to be a manager. Column 1 shows coefficients from a model where 'willingness to manage' (on a scale from 1-10) is regressed on a full set of individual characteristics. The three variables that are mostly strongly correlated with wanting to be in charge are extraversion, risk appetite, and being male. Columns 2 and 3 present regression results separately for men and women. The relationship between high extraversion and wanting to be a manager is driven largely by men.

Table 2: Associations with 'willingness to manage' ['how much do you want to be a manager, on a scale of 1 to 10']

	Regression Coefficients		
Dependent variable: preference to be in charge (scale of 1-10)	Full Sample (1)	Men (2)	Women (3)
Demographic Characteristics			
Female	-0.46		
	(0.27)		
Age (yrs)	0.00	0.01	-0.04
	(0.05)	(0.07)	(0.08)
Graduate student	0.26	0.61	-0.02
	(0.36)	(0.51)	(0.52)
Years of work experience	0.02	0.08	-0.04
	(0.06)	(0.08)	(0.08)
Skill Measures			
Production skills	0.27	0.43	0.03
	(0.15)	(0.19)	(0.02)
Economic decision making (AG)	0.21	0.60	-0.18
	(0.15)	(0.21)	(0.22)
Fluid IQ (CFIT)	0.23	0.04	0.42
	(0.15)	(0.21)	(0.22)
Emotional perceptiveness (RMET)	-0.13	-0.02	-0.17
	(0.15)	(0.21)	(0.22)
Personality and Working Styles			
Extraversion (Big5)	0.49	0.78	0.22
	(0.15)	(0.23)	(0.20)
Openness (Big5)	0.23	0.46	0.15
	(0.16)	(0.25)	(0.22)
Agreeableness (Big5)	-0.24	-0.50	-0.01
	(0.17)	(0.25)	(0.24)
Emotional stability (Big5)	0.33	0.26	0.34
	(0.15)	(0.23)	(0.21)
Conscientiousness (Big5)	-0.01	0.06	0.05
	(0.16)	(0.22)	(0.24)
Political Skill Inventory	0.09	0.06	0.20
	(0.16)	(0.23)	(0.24)
Risk appetite	0.42	0.43	0.39
	(0.15)	(0.21)	(0.21)
n	509	265	244

Notes: The dependent variable is each participant's willingness to be a manager on a scale from 1 to 10 'where 1 indicates a strong preference for NOT being a manager and 10 indicates a strong preference for being a manager'. The median response is 6 and the sd is 3. In eliciting participants' willingness to be a manager, we informed them that they should choose the role that they think 'best fits your skills'. The female variable is equal to 1 if participants identify as female, and 0 otherwise. The 'graduate student' variable is equal to one if participants were graduate students, and zero if they were undergraduates. All skill and personality variables have been standardized to be z-scores. Standard errors are in parentheses, calculated at the individual level. As per our pre-registered analysis plan, our goal here is to explore correlations, which is why we do not report a multiple-hypothesis-test correction (Parker and Weir, 2022b)

3 Identification strategy

3.1 Notation and setup

Let individuals be indexed by $i = 1, ..., n$. Individuals are randomly assigned to groups of three people, with groups indexed by g. Across the experiment there are n_g groups. Let I_{ig} be a binary indicator equal to one if participant i is in group g and zero otherwise. I_{ig} is governed by the randomization process described in Figure 2.

The experiment contains two roles: manager and worker. Each group contains one manager and two workers. Let M_{ig} be a binary indicator equal to one if participant i is the manager for group g and zero otherwise. Similarly, let W_{ig} be a binary indicator of whether participant i is a worker in group g.

Last, we have a set of variables that describe task performance. X_i measures individual productivity on the underlying tasks (i.e., scores on the numerical, spatial, and analytical tests completed by individuals before the group session). G_g denotes the performance of group g on the Collaborative Production Task.[14]

Some groups may perform well simply because they are randomly assigned participants with high levels of productive skill. To control for this, consider a simple model for the output of group g:

$$G_g = \gamma_M \sum_i X_i M_{ig} + \gamma_W \sum_i X_i W_{ig} + \epsilon_g \tag{1}$$
$$\epsilon_g \sim N(0, \sigma_G^2).$$

The terms $\sum_i X_i M_{ig}$ and $\sum_i X_i W_{ig}$ separately measure the productive skill of the manager and the workers in group g. By controlling for these measures separately we allow for the possibility that the productive skills of managers and workers differentially affect group output. The individual scores X_i come from the tests administered to participants before group testing.

3.2 Estimating manager performance

The residuals ϵ_g in equation (1) can be thought of as a measure of group performance, adjusted for random differences in each group's endowment of productive skill. If participants were only assigned to one group, it would be impossible to determine whether variation in ϵ_g arises from unmeasured individual attributes such as management skill,

[14]Group testing involves 4 rounds. In each round, groups face different questions. Following our analysis plan, we remove round effects by normalizing G_g scores within each round such that the distribution of scores within a round has mean 0 and standard deviation 1.

or from group dynamics between team members (Weidmann and Deming, 2021). However, by repeatedly randomly assigning managers to multiple groups, we can estimate the average causal impact that each manager has on group performance, after controlling for differences in productive skill:

$$\hat{a}_i = \frac{1}{\sum_g M_{ig}} \sum_g M_{ig} \hat{\epsilon}_g \tag{2}$$

In our framework, $\hat{a}_i$ is an estimate of the average manager's causal contribution, conditional on each group's endowment of task-specific skill. Because we only randomly assign managers to four teams, $\hat{a}_i$ is relatively noisy at the individual level. Thus, following our analysis plan, we focus on the question of whether α_i are correlated within managers—i.e., whether managers have a consistent impact on their teams, after controlling for productive skill. To do this, we fit a multilevel model:

$$\begin{aligned} \hat{\epsilon}_{gi} &= \alpha_i + e_{gi} \\ \alpha_i &\sim N(0, \sigma_\alpha^2) \\ e_{gi} &\sim N(0, \sigma^2) \end{aligned} \tag{3}$$

Our focus is σ_α, the standard deviation of the α_i estimates. In model (3), $\hat{\epsilon}_{gi}$ is a vector of skill-adjusted group performance ($1 \times n_g$), α_i is a random manager effect for individual i, and e_{gi} is residual error. The subscript i is included to indicate that this analysis examines variation at the level of individual managers. $\hat{\sigma}_\alpha$ is our estimate of the typical "manager effect": i.e., the impact on groups of having a manager who is 1 SD above average in terms of their managerial performance.

3.3 Estimating worker performance

Our framework analogously allows for the estimation of worker effects. To do this, we modify equation (2) to estimate the average causal effect that each worker has on their group, conditional on the group's endowment of productive skill:

$$\hat{\Omega}_i = \frac{1}{\sum_g W_{ig}} \sum_g W_{ig} \hat{\epsilon}_g \tag{4}$$

A similar substitution can be made to equation (3) to estimate the typical worker effect, defined as σ_Ω:

$$\begin{aligned}\hat{\epsilon}_{gi} &= \Omega_i + e_{gi} \\ \Omega_i &\sim N(0, \sigma^2_\Omega) \\ e_{gi} &\sim N(0, \sigma^2)\end{aligned} \tag{5}$$

In equation 5, $\hat{\epsilon}_{gi}$ is a vector of skill-adjusted group performance of length $(1 \times 2n_g)$, reflecting the fact that there are two workers in each group. Ω_i is a random worker effect for individual i on group g.

3.4 Inference

In our main analyses, we estimate the magnitude of the typical manager effect ($\hat{\sigma}_\alpha$) using equation (3). We compare this to a null hypothesis that managers have no impact on their teams after controlling for each team's endowment of productive skill. Given our somewhat complicated design, our preferred and pre-registered inferential approach is to calculate p-values using randomization inference. For robustness, we also report alternative estimates of uncertainty using a Wald estimator and Profile Likelihood estimates.[15]

The randomization inference procedure has four steps. First, we control for group differences in productive skill by estimating model (1). Second, we simulate five thousand random allocations of individuals to groups. These random allocations are blocked on 'experimental round' and 'role', so that in each simulated allocation, we observe every participant the same number of times – and in the same role – as we do in the experiment. Third, we fit models (2) and (3) for each simulation and estimate $\hat{\sigma}_{\alpha(NULL)}$. Fourth, we compare the observed manager effect $\hat{\sigma}^2_\alpha$ to the simulated distribution under the null and calculate the frequency with which draws from the null distribution are greater than $\hat{\sigma}^2_\alpha$, i.e., we estimate $Pr(\hat{\sigma}^2_{\alpha(NULL)} > \hat{\sigma}^2_\alpha)$. This is our p-value (Ernst, 2004).

4 Main Results

Our experimental design and identification strategy were pre-registered at the AEA RCT registry. Any deviations from this plan are noted in footnotes. This section presents our main pre-registered results. Section 5 presents exploratory analyses and evidence for mechanisms.

[15] The Wald estimator assumes a symmetric sampling distribution, which may not hold when estimating a variance parameter. Profile Likelihood confidence intervals are based on a chi-squared distribution and may be more suitable for a non-normal distribution bounded at zero (Venzon and Moolgavkar, 1988)

4.1 Can we identify good managers?

We estimate large and stable manager effects using our repeated random assignment method. The top row of Table 3 presents estimates of manager effects ($\hat{\sigma}_\alpha$), accompanied by p-values from each inference method. The second row of Table 3 presents analogous estimates for workers. The middle panel of the table presents coefficients on the measures of production skills used to condition group scores in model (1). These are included to contextualize the magnitude of manager effects. Standard errors for these coefficients are presented in parentheses.

Column 1 presents our pre-registered results. The typical manager effect is 0.23 standard deviations ($p = 0.04$). The coefficients on production skills—both for workers and managers—are positive and significant ($p < 0.001$) illustrating that a team's endowment of productive skill is strongly predictive of group success. The manager effect in column 1 is about 90% as large as the coefficient on workers' production skills as a group (around 0.26 sd).[16] This suggests that having a good manager matters about as much to group performance as the total productivity of both workers combined. In a formal test, we cannot reject the hypothesis that manager effects are equal in magnitude to production skills.

To illustrate the total average causal contribution managers and workers have on groups, Column 2 presents results without conditioning on production skills, so that model (1) is replaced with a null model. Removing the conditioning step increases the average manager effect from 0.23 sd to 0.29 sd and dramatically increases the magnitude of the worker effect, from 0.04 sd (ns) to 0.21 sd ($p = 0.03$). The elevated importance of worker effects in column 2 is not surprising given the nature of the Collaborative Production Task. Workers primarily contribute to group success through their ability to solve the problems to which they are assigned. When we condition on production skills, the worker effects decrease substantially.

Columns 3 to 6 present robustness checks. In Column 3 we control for whether managers knew or were friends with any of their team members outside the context of the experiment. This has a very small impact on the manager effect ($\hat{\sigma}_\alpha = 0.22$, p=0.05). In Column 4 we condition on each manager's risk appetite to control for the possibility that people who select into the manager role do so because there are sharper incentives. This leaves the results unchanged. Column 5 adds controls for the variance of individual production scores within a group, which again has no impact on estimates of manager effects.[17] Finally, in column 6 we condition group scores on a very fine-grained set of

[16]We compute the sum of both workers' production skills and then standardize this and all other skill measures to have mean = 0 and sd = 1, which allows us to compare magnitudes.

[17]For each group, we calculate the sd of the 9 individual measures of task skill (3 people with scores on 3 modules each). This is another approach to conditioning on production skills as, plausibly, the

skill measures, including separate controls for all sub-task measures (numerical; analytical; spatial) for both managers and workers. This has virtually no impact on manager effects ($\hat{\sigma}_\alpha = 0.22$, p=0.05), but reduces the worker effect to zero. As noted above, this is unsurprising as workers primarily contribute through their ability to solve problems.

To further assess the robustness of the worker and manager effects, we test whether estimates of α_i and Ω_i predict performance out-of-sample. To do this we perform a leave-one-out (LOO) procedure, in which we remove one of the four rounds of data, then calculate the average causal contributions of individual managers (α_i^{LOO}) and workers (Ω_i^{LOO}). We then assess whether these contributions predict whether a group will be successful in the holdout data.[18] We repeat this procedure for each of the four rounds and estimate the following model:

$$G_g = \beta_0 + \sum_i \hat{\alpha}_i^{LOO} M_{ig} + \sum_i \hat{\Omega}_i^{LOO} W_{ig} + \epsilon_g \tag{6}$$

The results are presented in Table 4. Columns 1 to 4 show the LOO analyses for each holdout round. These are noisier than our main analysis, as manager and worker effects are now based on only 3 random assignments. Column 5 aggregates the data and demonstrates that, on average, manager contributions predict out-of-sample group performance ($p < 0.01$). The point estimate for worker contributions is positive but less than half the magnitude of the manager association and not statistically significant. Overall, our LOO analysis suggests that the manager effects we are estimating robustly predict performance.

more skill variance a team has, the more chance they have to specialize and do well in the task.

[18] We follow our pre-registered approach, with one necessary deviation. Our intention was to use manager and worker effects from analyses that conditioned on production skills (i.e., the analysis presented in column 1 in Table 3). However, as conditioning on production skills often makes it impossible to estimate worker effects in small samples, we instead examine the total contribution that workers and managers typically make to groups, i.e., $\hat{\alpha}_i^{LOO}$ and $\hat{\Omega}_i^{LOO}$ are calculated using the approach outlined in column (2) of Table 3.

Table 3: Estimating the magnitude of manager and worker effects

	Dependent variable: Group Performance (G)					
	(1)	(2)	(3)	(4)	(5)	(6)
Manager effect ($\hat{\sigma}_\alpha$)	0.228	0.286	0.219	0.218	0.220	0.218
[Randomization inference]	[0.04]	[<0.01]	[0.05]	[0.05]	[0.04]	[0.05]
{Wald}	{< 0.005}	{< 0.005}	{< 0.005}	{< 0.005}	{< 0.005}	{< 0.005}
(Profile likelihood)	(0.04)	(0.01)	(0.05)	(0.05)	(0.04)	(0.05)
Worker effect ($\hat{\sigma}_\Omega$)	0.041	0.207	0.078	0.066	0.051	0
[Randomization inference]	[0.48]	[0.03]	[0.39]	[0.41]	[0.46]	
{Wald}	{0.40}	{< 0.005}	{0.04}	{0.04}	{0.20}	
(Profile likelihood)	(0.49)	(0.03)	(0.40)	(0.44)	(0.47)	
Controls						
Manager's production skills[1]	0.208		0.215	0.209	0.198	
	(0.040)		(0.040)	(0.041)	(0.041)	
Workers' production skills[2]	0.261		0.263	0.263	0.238	
	(0.035)		(0.035)	(0.035)	(0.040)	
Manager familiar w/ participants[3]			x	x	x	x
Manager risk appetite[4]				x	x	x
Variance team production skills[5]					x	x
Granular production skills[6]						x
Counts						
Groups [4 rounds per person]	728	728	700	700	700	700
Managers	186	186	176	176	176	176
Workers	369	369	357	357	357	357

Notes: the dependent variable is G_g. All models include fixed effects for whether the session appointed managers through a lottery or via self-selection. Manager and worker effects are estimated using model (3). We report p-values using three different approaches to inference, in all cases the null hypothesis being tested is that $\hat{\sigma}_x = 0$. [1]Manager production skills are defined at the individual level as the standardized score across the numerical, analytical and spatial tasks (sd=1, mean=0). [2]Worker production skills are defined at the group level as the mean score of both workers for the numerical, analytical and spatial tasks (and standardized so that this variable has sd=1 and mean=0 across all groups). [3]Familiarity with other participants is a binary variable, based on whether any of the managers reported being friends with, or knowing, any of the workers in their groups. [4]Manager risk appetite is measured on a scale of 1-10. [5]For each group, we calculate the variance of task skill within the team, which includes 9 separate measures of skill (3 people, with 3 measures each). [6]Granular task skills are the scores on the numerical, analytical and spatial tasks. In this specification, model (1) includes 3 covariates for manager skills (numerical manager; analytic manager; spatial manager) and 3 for the average of the workers. The estimate of 0 for the worker effect comes from our multilevel model (3), which estimates zero variance at the level of individual workers.

Table 4: Predicting manager and worker contributions out-of-sample

	Group performance in hold out round				
	Round 1	**Round 2**	**Round 3**	**Round 4**	**Overall**
	(1)	(2)	(3)	(4)	(5)
Manager contribution LOO ($\hat{\alpha}_i^{LOO}$)	0.388	0.237	0.185	0.385	0.323
	(0.147)	(0.179)	(0.212)	(0.181)	(0.105)
Worker contribution LOO ($\hat{\Omega}_i^{LOO}$)	-0.001	—	0.261	0.131	0.130
	(0.147)		(0.212)	(0.181)	(0.105)
Observations	186	182	180	180	546
R^2	0.037	0.010	0.013	0.030	0.021
Adjusted R^2	0.026	0.004	0.002	0.019	0.017

Notes: The dependent variable is group score in the holdout round of data. 'Manager contribution LOO' is defined using the remaining 3 rounds of data. The same is true for worker contributions. There is no estimate for worker contribution when round 2 data is held out, as the estimate for σ_Ω in this case is 0, meaning we cannot estimate worker effects. Standard errors are in parentheses and are calculated at the group level.

4.2 What characteristics predict being a good manager?

Having causally identified the contribution that managers make, we ask: *what are the characteristics of good managers?* Table 5 explores the correlates of manager contributions beyond being productive at the underlying tasks.[19] Table 5 separately reports predictors of management performance for lottery managers (column 1) as well as the sample of self-promoted managers (column 2). We focus primarily on the lottery managers, as the relationships between manager performance and broad skills/traits in the self-promoted arm are moderated by the filter of self-promotion. Only two measures predict manager performance in the lottery arm: fluid intelligence (CFIT) and economic decision-making (scores on the Assignment Game), both of which are statistically significant at the less than one percent level.[20]

Among self-promoted managers (column 2) we find negative correlations between management performance and both self-reported extraversion ($\rho = -0.24$, $p < 0.05$) and self-reported political skill ($\rho = -0.26$, $p < 0.05$). In other words, when managers are selected by self-promotion, the managers who think they are a "people person" are less successful in the job.

Table 6 shows that the two reliable predictors of management performance - economic decision-making skill and fluid intelligence - are robust to a wide range of controls, including demographics, education and work experience, and measures of emotional perceptiveness and personality.

[19]We focus on contributions measured from our preregistered model.
[20]It is useful to note that these correlations are explanatory (Parker and Weir, 2022a).

Table 5: Correlates of management performance

Correlation with manager contributions Correlation with $\hat{a}_i$: This is a measure of manager contribution, conditional on the team's endowment of production, as per our pre-specified model[a]			
	Lottery managers n=90 (1)	**Self-promoted managers** n=96 (2)	**Full sample** n=186 (3)
Skill assessments			
Fluid intelligence (CFIT)	0.24	0.31	0.27
	(0.02)	(<0.01)	(<0.01)
Economic Decision-making (AG)	0.24	0.10	0.16
	(0.02)	(0.34)	(0.03)
Emotional perceptiveness RMET)	-0.02	0.20	0.09
	(0.83)	(0.05)	(0.21)
Personality and working styles			
Extraversion (Big5)	-0.06	-0.24	-0.15
	(0.57)	(0.02)	(0.05)
Openness (Big5)	-0.17	-0.05	-0.11
	(0.12)	(0.64)	(0.13)
Agreeableness (Big5)	-0.18	0.02	-0.06
	(0.10)	(0.87)	(0.39)
Neuroticism (Big5)	0.12	0.04	0.06
	(0.28)	(0.74)	(0.40)
Conscientiousness (Big5)	0.00	-0.04	-0.02
	(0.97)	(0.67)	(0.74)
Political Savvy (PSI)	-0.03	-0.26	-0.15
	(0.76)	(0.01)	(0.04)
Risk appetite	0.06	-0.14	-0.05
	(0.57)	(0.18)	(0.53)
Demographics			
Age	-0.01	0.16	0.07
	(0.95)	(0.14)	(0.36)
Female	-0.13	-0.13	-0.12
	(0.23)	(0.22)	(0.12)
Years of work experience	-0.08	0.05	-0.01
	(0.48)	(0.62)	(0.88)

Notes: [a]The measure comes from the analysis reported in column (1) of Table 1. To express uncertainty, p-values are in parentheses. As per our pre-registered analysis plan, our goal here is to explore correlations, which is why we do not report a multiple-hypothesis-test correction (Parker and Weir, 2022b)

Table 6: Robustness of relationship between skill assessments and management (lottery arm)

	Dependent variable: management contribution, $\hat{a}$ (pre-specified model)											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Economic decision-making	0.298	0.348	0.389	0.365	0.405						0.245	
(Assignment Game)	(0.131)	(0.139)	(0.149)	(0.150)	(0.157)						(0.182)	
Fluid intelligence						0.225	0.235	0.301	0.321	0.313	0.220	
(CFIT)						(0.098)	(0.100)	(0.104)	(0.104)	(0.112)	(0.131)	
Combined Decision making												0.365
(AG+CFIT)												(0.117)
RMET		x	x	x	x		x	x	x	x	x	x
Demographics			x	x	x			x	x	x	x	x
Education and work				x	x				x	x	x	x
Personality					x					x	x	x
Obs	90	90	86	86	86	90	90	86	86	86	86	86
Adjusted R^2	0.045	0.047	0.043	0.043	0.045	0.046	0.039	0.060	0.084	0.059	0.070	0.083

Notes: The dependent variable in these regressions is the manager contributions from our pre-specified model (that controls for production skills). We focus on data from the random arm of the experiment (n=90). 4 participants have post-surveys missing, which reduces the sample to 86 for some specifications. Demographics includes age, gender and ethnicity. Education and work includes years of work experience, and the highest level of education; personality is the five Big5 measures. The 'Combined Decision Making' variable is a simple average of a participant's standardized score on the Assignment Game and the CFIT test.

4.3 Self-promoted managers perform worse than randomly assigned managers

Next, we examine the performance of participants who express a strong desire to be managers. Because we randomly assign the managerial selection rule, we can cleanly compare manager performance across treatment arms to learn whether people who self-promote into management perform better than randomly assigned managers.

Table 7 contrasts the performance of groups managed by a 'lottery manager' with those of a 'self-promoted manager'. We regress group performance G_g on the random indicator of 'self-promotion' (equal to 1 if managers were in the self-promotion arm). Column (1) shows that self-promoted managers are on average *worse* than lottery counterparts, although the estimate is only statistically significant at the ten percent level (0.13sd, $p = 0.09$). Columns 2 through 8 add a series of controls for group characteristics, including group endowments of IQ, emotional perceptiveness, and productive skill. We also include controls for manager risk appetite and demographic factors such as personality and levels of education. Throughout these specifications, the magnitude of the difference between self-promoted and lottery managers is around -0.10 standard deviations, and on the margin of statistical significance. However, the magnitude is relatively large. Groups with self-promoted managers perform about as poorly on average as groups with fluid intelligence that is one full standard deviation below average.

Why might people who strongly prefer to be managers perform worse than those who got the job by chance? We hypothesize that self-promoting managers are overconfident. We test this with an exploratory analysis in which we asked people to reflect on whether they thought they were "much worse," "worse," "average," "better," or "much better" than their peers. We then computed an individual-level measure of overconfidence by regressing self-reported performance on each person's causal contribution to the team (α for managers, Ω for workers) and capturing the residual. People who want to be in charge are much more overconfident that people who do not have strong preferences for being a manager ($d = 0.41$ sd, $p < 0.01$). This reflects the results from field studies on the overconfidence of managers and executives (e.g. Malmendier and Tate (2015)).

Another reason for the poor performance of self-promoted managers may be social skills. This could be directly related to overconfidence, because overrating one's own abilities may lead to being less attentive to the emotions and capacities of others. Notably, we find that managers' overconfidence is strongly negatively related to their emotional perceptiveness as measured by the RMET (correlation $= -0.33$, $p < 0.001$).

We also find a strong negative correlation between self-reported people skills and managers' performance on RMET in the self-promotion arm (correlation $= -0.37$, $p <$

0.001).[21] However, the relationship was not significant for managers in the lottery arm. This is consistent with Heck et al. (2024), who find a strong negative relationship between self-reported social intelligence and skill-based assessments of social intelligence. Overall, the evidence suggests that the poor performance of self-promoted managers is driven by their overconfidence, especially in terms of their social skills.

4.4 Quantifying the impact of manager selection mechanisms

We quantify the benefits of skill-based promotions by explicitly comparing the impact that different selection mechanisms have on average manager contributions. In addition to comparing self-promoted and lottery managers, we use data from the lottery arm to simulate counterfactuals in which managers are selected on specific skills.

As an example, consider selecting managers based on the Peter Principle (Peter and Hull, 1969). In our case, this would mean ranking participants in terms of their individual production skills and appointing the top 1/3rd of participants as managers (since 1 in 3 people in the experiment are managers). We can estimate the average quality of managers under the Peter Principle by ranking the top 1/3rd of managers in the lottery arm and calculating their average manager performance.[22] Analogously, we can look at any other individual skill as a basis for manager selection.

Figure 6 compares the results of six selection mechanisms. We examine self-promotion, lottery, and choosing managers based on four specific skills: fluid IQ (CFIT scores), economic decision-making (AG scores), emotional perceptiveness (RMET scores), and individual production skills (i.e. the Peter Principle).[23] Different selection mechanisms have large impacts on manager quality and group performance. Compared to a regime of self-promotion, selecting managers based on economic decision-making skills yields managers who are 0.6 standard deviations better in terms of their manager effects. Translating this into group performance, this is equivalent to replacing an average worker in every group with a worker in the 99th percentile in terms of productivity. This suggests that organizations would likely benefit from considering a wide pool of potential managers, not just people who are proactive in seeking management roles.

[21] In this exploratory analysis, we calculate self-reported people skills as an average of Big 5 extraversion, and scores on the Political Skill Inventory (PSI).

[22] It is important to use the lottery arm rather than the self-promotion arm, because self-selection is correlated with skills and other characteristics. For example, you may also be selecting for extraversion. This undermines the ability to estimate the manager quality under the Peter Principle (or based on other skills).

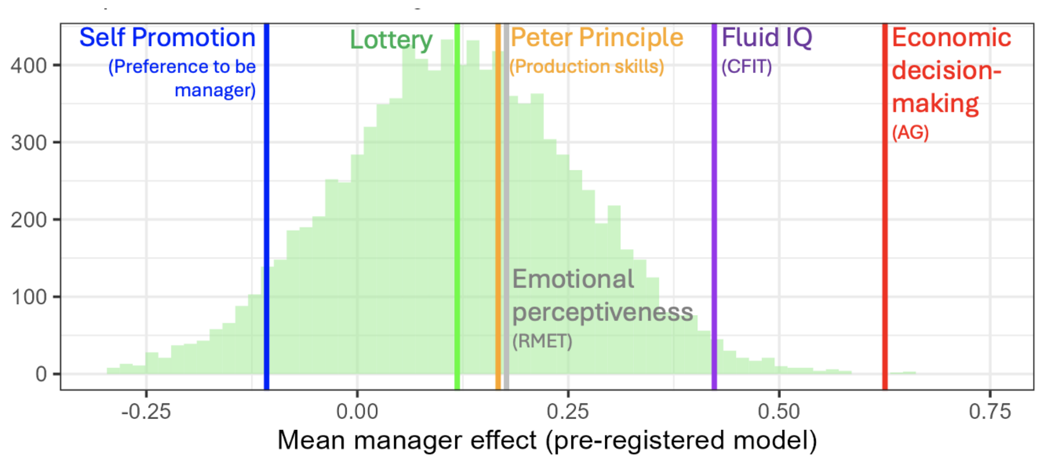
[23] It is useful to acknowledge that skills such as IQ are not randomly assigned.

Table 7: Difference in performance of team led by 'lottery manager' vs 'self-promoted manager'

Dependent var: group scores (G_g)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Self-promoted vs lottery managers	**-0.125**	**-0.124**	**-0.115**	**-0.111**	**-0.101**	**-0.098**	**-0.094**	**-0.096**
(standard errors)	**(0.074)**	**(0.070)**	**(0.069)**	**(0.069)**	**(0.069)**	**(0.070)**	**(0.070)**	**(0.070)**
Production skills[1]		0.185	0.138	0.134	0.130	0.131	0.131	0.130
Fluid IQ (CFIT)[2]			0.114	0.108	0.104	0.113	0.113	0.112
Economic Decision-Making (AG)[2]				0.019	0.011	0.010	0.010	0.009
Emotional Perceptiveness (RMET)[2]					0.028	0.015	0.019	0.023
Risk appetite[3]						-0.005	-0.003	-0.003
Know others in experiment[4]							0.074	0.071
Education and work experience[5]								x
Obs	728	728	728	728	728	700	700	700
Adjusted R^2	0.003	0.101	0.134	0.134	0.135	0.141	0.141	0.139

Notes: The dependent variable is group performance (G_g). Standard errors are presented under the main coefficients in parentheses and are clustered at the group level. [1]Production skills are defined at the group level as the mean score (averaging across group members) on the individual tests of numeracy, spatial, and analytical reasoning and standardized to have mean=0 and sd=1 across groups.. [2]Measure is defined at the group level as the mean score (averaging across group members) of an individual measure, e.g., the Fluid IQ test CFIT (standardized to have mean=0 and sd=1). [3]Risk appetite is a self-reported measure of the risk tolerance of the group's manager, on a scale of 1-10. [4]Know others in experiment is a binary indicator at the group level equal to one if the manager knew either of the workers. [5]Education and work experience" represent two group-level variables: education is the percent of group members who have graduated from their undergraduate program, work experience is the group mean number of years of work experience.

Figure 6: Comparing methods of appointing managers



Notes: To demonstrate sampling uncertainty, Figure 6 includes a sampling distribution from the random arm of the experiment. Each draw represents the mean score from a subset of 30 managers from the lottery arm (n=30 was chosen as this matches the size of the subsets defined by the other regimes, e.g. the Peter Principle where we select top tercile of managers based on production skills. Note that the mean manager effect in the lottery arm is not zero, as manager effects are normalized across the whole sample (which includes self-promoted managers, who typically perform worse than their lottery counterparts). Note too that the difference between average manager quality for AG and CFIT is not statistically significant. .

5 Mechanisms

In this section we examine three potential mechanisms for improving performance – monitoring, allocation according to comparative advantage, and motivating workers to exert effort.

5.1 Monitoring

The first mechanism we examine is the extent to which managers monitor team members. Specifically, we argue that active monitoring helps avoid situations where workers are wasting their time on unhelpful tasks. To measure this, we focus on the task that workers are assigned to as each group's time expires. Recall that our weakest-link scoring rule means that it's the minimum module score that defines group scores G_g. The scoring rule is emphasized in the instructions and is the subject of specific practice questions before the group session begins. If, when the time runs out, any team member is working on a module whose score is substantially greater than the minimum (which determines the final score), that person is effectively wasting their time because their effort will not increase the group's score.

We arbitrarily define a module score as being "substantially greater" than the team score if it is 10 points higher than the module score, although our analysis is not sensitive to this particular threshold. We define a monitoring failure by the manager as having any group member working on a module at the end of the Collaborative Production Task that is substantially greater - e.g. 10 points higher - than the minimum module score.

Over the course of the experiment, 16% of groups ($n = 728$) finish the task with at least one person working on a module that was not contributing to group success. Failures in monitoring were strongly related to overall manager contributions. The bivariate correlation between monitoring errors and manager performance is -0.40 ($p < 0.001$, $n = 186$). A manager 1sd above average reduced the error rate from 16% to 8%. In other words, good managers had half the rate of monitoring errors.

5.2 Allocation According to Comparative Advantage

Next, we examine the quality of managers' allocation decisions. To simplify the analysis, we focus on each manager's initial allocation, and we limit our analysis to groups who initially assigned one person to each of the three modules. More than 90% of groups used this strategy. Focusing on these decisions allows us to study whether initial allocations were optimal. Once the task begins, dynamics within the task make it difficult to cleanly identify optimal allocations.

With three people and three modules, there are six possible one-to-one initial assignments. We use information on participants' individual module test scores (assessed before the group session began) to assess whether or not managers found the optimal assignment. A group is considered optimally assigned if each participant is allocated to the module where they have a comparative advantage based on their individual scores.

The probability a manager finds the optimal starting assignment is positively associated with their manager performance ($\rho = 0.19$, $p < 0.01$). To quantify the impact on group performance, we compare groups whose managers always start with the optimum assignment ($n = 74$) with groups whose managers never start optimally ($n = 42$). Groups whose managers always start with the best assignment scored 0.52sd higher than groups with managers who never start with the best allocation ($p < 0.01$). This suggests that figuring out the best allocation of workers to tasks is a strong component of management performance.[24]

5.3 Motivating Workers to Exert Effort

The final mechanism we examine is worker motivation. The group task involves three periods of intensive problem solving, each of which lasts two minutes.[25] The tasks are repetitive and cognitively demanding. Since workers do not have financial incentives to exert effort, they may lose motivation over the course of the experiment.

To test whether some managers do a better job of maintaining motivation, we partition group performance into 3 parts, reflecting the three two-minute problem-solving periods. This partitioning corresponds to the natural breakpoints of each session, since managers are given 60 seconds to refocus and motivate their team between each two-minute problem-solving period.

We estimate manager contributions separately in each two-minute period by estimating:

$$\alpha_i = \gamma_1 \alpha_i^{first2} + \gamma_2 \alpha_i^{middle2} + \gamma_3 \alpha_i^{last2} + \epsilon_i$$

[24]A further related role of the manager is to develop strategies during the session. To measure this we use a post experiment survey question: "As a manager did you have a strategy? And if so what was it?" One issue with this measure is that it is based on the manager's post-experiment rationalization rather than actual task behavior, which might be influenced by factors such as confidence, with more confident managers articulating strategies clearly; and some descriptions not fully capturing the nuances of strategic thinking. Nevertheless, it serves as a useful test of the perceived complexity of the manager's strategy. We create two measures from this data. First, we create a variable that is equal to the number of characters written in the strategy description. Second, we asked GPT4 to code strategies as either good or bad. To train the AI we gave GPT4 three examples of a sophisticated strategy and three examples of a bad strategy. Both measures have very similar predictive power in terms of manager performance (0.23, p=0.01 for the first measure and 0.21, p=0.03 for the second). This provides further evidence that manager strategisation is important for team outcomes.

[25]These 2-minute problem solving periods are divided by 1 minute breaks.

where: α_i is the manager contribution of participant i estimated across the full experiment using our pre-registered model. α_i^{period} is the manager contribution of participant i estimated using data that only includes team performance during period $x \in$ (first 2 minutes, middle 2 minutes, last 2 minutes).

The results are presented in Table 8. Columns 1-3 show results separately for each two-minute period, while Column 4 puts them all together in the same regression. Overall manager performance is mostly strongly influenced by the final two minutes. Column 4 shows that group performance in the last two-minute period is about 50% more influential than performance in the first two minutes (p=0.038). Translating these coefficients into raw output, teams led by a manager who is 1 SD above average solve 0.6 more problems in the final two minutes but only 0.3 problems more in the first two minutes. We interpret the differences in the importance of the first and last two minutes of the experiment as predominantly being driven by motivation and engagement. [26]

Table 8: What period is matters most for manager performance?

	Dependent variable: α_i			
	(1)	(2)	(3)	(4)
Start	0.439			0.376
	(0.077)			(0.048)
Middle		0.517		0.429
		(0.074)		(0.048)
End			0.668	0.529
			(0.065)	(0.049)
Constant	0.028	0.028	0.028	0.028
	(0.077)	(0.074)	(0.065)	(0.047)
Obs	147	147	147	147
R^2	0.183	0.253	0.423	0.701
Adjusted R^2	0.177	0.248	0.419	0.694

Notes: the dependent variable is α_i, estimated using the pre-registered model. The covariates are alphas estimated using 1/3 of the data: 'start' represents alphas when we use team scores captured after the first third of the experiment; 'middle' correspond to alphas estimated using team scores from only the period between the first and second breaks; 'end' is alphas from the team performance on the last 2 minutes of the experiment. Note that we only have data for 147 out of the 186 managers. Other managers did not submit allocations in a way that allowed us to capture their teams score at the breakpoints.

[26] Another potential explanation is that people improved as time went on - however across the four rounds of the experiment we find no secular improvement in scores, suggesting a lack of learning effects.

6 Conclusion

This paper develops an experimental methodology for prospectively identifying good managers. We repeatedly randomly assign managers to different groups of workers who perform a Collaborative Production Task together, and in each case, we estimate the contribution of the manager to group performance after conditioning on workers' skills. Over multiple random assignments, some managers consistently cause their teams to exceed predicted performance. Good managers are roughly twice as valuable as good workers, consistent with studies of managerial performance in other settings. Using our novel method we estimate the predictors of management performance. Good managers have higher fluid intelligence and score higher on a test of economic decision-making skill. We find no difference in average managerial performance by gender, age or ethnicity.

We also find that self-promoted managers perform worse than managers who are randomly assigned to the role. We show that this is likely due to overconfidence. Managers whose actual performance is worse than their self-reported performance have stronger preferences to be in charge. Self-promoted managers have higher reported social skills, but do worse on a widely-used skill-based test of emotional perceptiveness. We also show that self-nomination is highly correlated with extraversion and self-reported people skills, especially among men. Finally, we show that good managers increase group performance by monitoring workers to avoid wasting time, by allocating workers to tasks that maximize their comparative advantage, and by motivating them to exert effort.

Overall, we find that good managers matter, and that skills are much better predictors of manager performance than personality traits or preferences. This is important because preferences for leadership and qualities like extraversion and self-confidence greatly increase the probability of promotion in many workplaces. Our results suggest that selecting managers based on economic decision-making skill rather than self-promotion would increase organizational performance. Our findings also suggest that firms who screen for skills such as economic decision making could see improvements in team performance.[27] Our method can be used easily by firms and organizations seeking to hire and promote effective managers.

[27] This is consistent with results from field settings which find an association between manager cognitive skill and survey based measures of productivity (Adhvaryu et al., 2022)

References

Adhvaryu, A., P. Gaule, and A. Nyshadham (2023): "Performance Pay and Worker Experience," The Quarterly Journal of Economics, 138, 511–560.

Adhvaryu, A., A. Nyshadham, and J. Tamayo (2022): "Managerial Quality and Productivity Dynamics," The Review of Economic Studies, 90, 1569–1607.

Ahearn, K., G. Ferris, W. Hochwarter, C. Douglas, and A. Ammeter (2004): "Leader political skill and team performance," Journal of Management, 30, 309–327.

Alchian, A. A. and H. Demsetz (1972): "Production, Information Costs, and Economic Organization," The American Economic Review, 62, 777–795.

Baron-Cohen, S., S. Wheelwright, J. Hill, Y. Raste, and I. Plumb (2001): "The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism," Journal of Child Psychology and Psychiatry, 42, 241–251.

Becker, A., T. Deckers, T. Dohmen, A. Falk, and F. Kosse (2012): "The relationship between economic preferences and psychological personality measures," Annual Review of Economics, 4, 453–478.

Bell, S. (2007): "Deep-Level Composition Variables as Predictors of Team Performance," The Journal of applied psychology, 92, 595–615.

Bennedsen, M., F. Perez-Gonzalez, and D. Wolfenzon (2007): "Do CEOs matter?" Tech. Rep. 2006-21, Center for Economic Institutions, Institute of Economic Research.

Benson, A., D. Li, and K. Shue (2019): "Promotions and the Peter Principle*," The Quarterly Journal of Economics, 134, 2085–2134.

Bhalotra, S., I. Clots-Figueras, L. Iyer, and J. Vecci (2022): "Leader Identity and Coordination," Review of Economics and Statistics, 105, 175–189.

Bloom, N., B. Eifert, A. Mahajan, D. McKenzie, and J. Roberts (2013): "Does Management Matter? Evidence from India," The Quarterly Journal of Economics, 128, 1–51.

Bloom, N., R. Lemos, R. Sadun, D. Scur, and J. Van Reenen (2014): "JEEA-FBBVA Lecture 2013: The New Empirical Economics of Management," Journal of the European Economic Association, 12, 835–876.

Bloom, N., R. Sadun, and J. Van Reenen (2016): "Management as a technology?" National Bureau of Economic Research.

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

Born, A., E. Ranehill, and A. Sandberg (2022): "Gender and Willingness to Lead: Does the Gender Composition of Teams Matter?" The Review of Economics and Statistics, 104, 259–275.

Brandts, J. and D. J. Cooper (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, 1223–1268.

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

Camerer, C. F. and R. A. Weber (2013): "Experimental Organizational Economics," in Handbook of Organizational Economics, ed. by R. Gibbons and J. Roberts, Princeton and Oxford: Princeton University Press, 153–215.

Caplin, A., D. J. Deming, S. Leth-Petersen, and B. Weidmann (2024): "Economic Decision-Making Skill Predicts Income in Two Countries," Working Paper 31674, National Bureau of Economic Research.

Chabris, C. F. (2007): "Cognitive and Neurobiological Mechanisms of the Law of General Intelligence," in Integrating the Mind: Domain General vs Domain Specific Processes in Higher Cognition, ed. by M. J. Roberts, Psychology Press, 449–491.

Chakraborty, P. and D. Serra (2023): "Gender and Leadership in Organisations: the Threat of Backlash," The Economic Journal, 134, 1401–1430.

Chamorro-Premuzic, T. (2019): Why do so many incompetent men become leaders?: (And how to fix it), Harvard Business Press.

Charness, G. and P. Kuhn (2011): "Lab Labor: What Can Labor Economists Learn from the Lab?" Elsevier, vol. 4A, chap. 03, 229–330, 1 ed.

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

——— (2020b): "Fool Me Once: An Experiment on Credibility and Leadership," The Economic Journal, 130, 2105–2133.

Desseranno, E., M. Stryjan, and M. Sulaiman (2019): "Leader selection service delivery in community groups," American Economic Journal: Applied Economics, 11, 240–267.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. Wagner (2011): "Individual risk attitudes: Measurement, determinants, and behavioral consequences," Journal of the European Economic Association, 9, 522–550.

Englmaier, F., S. Grimm, D. Grothe, D. Schindler, and S. Schudy (2024): "The Effect of Incentives in Non-Routine Analytical Team Tasks," Journal of Political Economy, 0, null.

Ernst, M. D. (2004): "Permutation methods: a basis for exact inference," Statistical Science, 676–685.

Ertac, S. and M. Y. Gürdal (2012): "Personality, Group Decision Making, and Leadership," Koç University-TUSIAD Economic Research Forum Working Papers.

Falk, A. and J. J. Heckman (2009): "Lab Experiments Are a Major Source of Knowledge in the Social Sciences," Science, 326, 535–538.

Feld, J., E. Ip, A. Leibbrandt, and J. Vecci (2022): "Identifying and Overcoming Gender Barriers in Tech: A Field Experiment on Inaccurate Statistical Discrimination," Tech. Rep. 9970, CESifo Working Paper.

Fenizia, A. (2022): "Managers and Productivity in the Public Sector," Econometrica, 90, 1063–1084.

Ferris, G., D. Treadway, R. Kolodinsky, W. Hochwarter, C. C. Kacmar, C. Douglas, and D. Frink (2005): "Development and Validation of the Political Skill Inventory," Journal of Management - J MANAGE, 31, 126–152.

Gosling, S., P. Rentfrow, and W. Swann (2003): "A Very Brief Measure of the Big-Five Personality Domains," Journal of Research in Personality, 37, 504–528.

Güth, W., M. V. Levati, M. Sutter, and E. van der Heijden (2007a): "Leadership and Cooperation in Public Goods Experiments," Journal of Public Economics, 91, 1023–1042.

——— (2007b): "Leadership and Cooperation in Public Goods Experiments," Journal of Public Economics, 91, 1023–1042.

Gächter, S., C. Starmer, and F. Tufano (2023): "Measuring "Group Cohesion" to Reveal the Power of Social Relationships in Team Production," The Review of Economics and Statistics, 1–45.

Haier, R., R. Colom, D. Schroeder, C. Condon, C. Tang, E. Eaves, and K. Head (2009): "Gray matter and intelligence factors: Is there a neuro-g?" Intelligence, 37, 136–144.

Hardeweg, B., L. Menkhoff, and H. Waibel (2013): "Experimentally Validated Survey Evidence on Individual Risk Attitudes in Rural Thailand," Economic Development and Cultural Change, 61, 859–888.

Heck, P. R., A. K. Brown, and C. F. Chabris (2024): "The Social Sensing Hypothesis," Manuscript.

Herbst, D. Z. and A. Mas (2015): "Peer effects on worker output in the laboratory generalize to the field," Science, 350, 545 – 549.

Hirshleifer, J. (1983): "From Weakest-Link to Best-Shot: The Voluntary Provision of Public Goods," Public Choice, 41, 371–386.

Hjort, J. and D. Chandler (2022): "How to Make the Hybrid Workforce Model Work," MIT Sloan Management Review, 63, 1–10.

Hoffman, M., L. B. Kahn, and D. Li (2017): "Discretion in Hiring*," The Quarterly Journal of Economics, 133, 765–800.

Hoffman, M. and S. Tadelis (2021): "The Performance Effects of Hiring Based on Referrals," National Bureau of Economic Research.

Hurtz, G. and J. Donovan (2000): "Personality and Job Performance: The Big Five Revisited," The Journal of applied psychology, 85, 869–79.

Kahneman, D. and G. Klein (2009): "Conditions for intuitive expertise: a failure to disagree," American Psychologist, 64, 515.

Kaplan, S. (2012): "Personality and the economics of personality," Journal of Economic Perspectives, 26, 133–154.

Larson, J. (2013): "In Search of Synergy in Small Group Performance," In Search of Synergy: In Small Group Performance, 1–427.

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

Malmendier, U. and G. Tate (2015): "Behavioral CEOs: the role of managerial overconfidence," Journal of Economic Perspectives, 29, 37–60.

Metcalfe, R., A. Sollaci, and C. Syverson (2023): "Managers and Productivity in Retail," .

Minni, V. (2023): "Making the invisible hand visible: Managers and the allocation of workers to jobs," POID Working Papers 080, Centre for Economic Performance, LSE.

Nevicka, B., F. S. Ten Velden, A. H. B. De Hoogh, and A. E. M. Van Vianen (2011): "Reality at Odds With Perceptions: Narcissistic Leaders and Group Performance," Psychological Science, 22.

Parker, R. and C. Weir (2022a): "Multiple secondary outcome analyses: precise interpretation is important," Trials, 23.

Parker, R. A. and C. J. Weir (2022b): "Multiple secondary outcome analyses: precise interpretation is important," Trials, 23, 27.

Peter, L. J. and R. Hull (1969): The Peter Principle: Why Things Always Go Wrong, New York: William Morrow and Company.

Potters, J., M. Sefton, and L. Vesterlund (2007): "Leading-by-example and signaling in voluntary contribution games: an experimental study," Economic Theory, 33, 169–182.

Queiro, J. (2022): "Good Bosses," Journal of Political Economy, 130, 489–529.

Reuben, E., P. Rey-Biel, P. Sapienza, and L. Zingales (2010): "The Emergence of Male Leadership in Competitive Environments," Journal of Economic Behavior & Organization, 83.

Riedl, A., I. M. T. Rohde, and M. Strobel (2015): "Efficient Coordination in Weakest-Link Games," The Review of Economic Studies, 83, 737–767.

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

Venzon, D. J. and S. H. Moolgavkar (1988): "A Method for Computing Profile-Likelihood-Based Confidence Intervals," Journal of the Royal Statistical Society. Series C (Applied Statistics), 37, 87–94.

Weidmann, B. and D. J. Deming (2021): "Team Players: How Social Skills Improve Team Performance," Econometrica, 89, 2637–2657.