

Ben Weidmann  
Joseph Vecci  
Farah Said  
Sonia Bhalotra  
Achyuta Adhvaryu  
Anant Nyshadham  
Jorge Tamayo  
David Deming

25/37

## Working paper

# How do you identify a good manager?

# How Do You Identify a Good Manager?

Ben Weidmann\*    Joseph Vecchi<sup>†</sup>    Farah Said<sup>‡</sup>    Sonia Bhalotra<sup>§</sup>  
Achyuta Adhvaryu<sup>¶</sup>    Anant Nyshadham<sup>||</sup>    Jorge Tamayo\*\*  
David Deming<sup>††</sup>

July 13, 2025

## Abstract

This paper introduces and validates a novel approach to measuring management skills. In a large pre-registered lab experiment we causally identify managerial contributions by randomly assigning managers to multiple teams and controlling for differences in individual skill. We find that manager contributions matter greatly for team success, and that people who want to be in charge perform worse than randomly assigned managers. Managerial performance is strongly predicted by economic decision-making skill, but not by demographic characteristics. LinkedIn data show that participants who succeed in the lab are substantially more likely to receive real-world promotions. We also measure the skills of store managers in a large retail firm and find that they predict store sales and other correlates of productivity, which aligns with our experimental results. A one standard deviation increase in manager quality increases annual per-store sales by \$4.1 million USD (25% increase). Selecting managers on skills rather than demographic characteristics or the desire to lead could substantially improve organizational performance.

*Keywords:* Management, Teamwork, Skills, Measurement, Experiment JEL Codes: M54, J24, C90, C92

---

\*Harvard Kennedy School and UCL benweidmann@hks.harvard.edu

<sup>†</sup>Department of Economics, University of Gothenburg. Joseph.Vecchi@economics.gu.se

<sup>‡</sup>Department of Economics, Lahore University of Management Sciences. farah\_said@lums.edu.pk

<sup>§</sup>University of Warwick, IFS, CESifo, CEPR, IZA. Sonia.Bhalotra@warwick.ac.uk

<sup>¶</sup>University of California San Diego, NBER, J-PAL, BREAD, Good Business Lab; aad-hvaryu@ucsd.edu

<sup>||</sup>University of Michigan, NBER, J-PAL, BREAD, Good Business Lab; nyshadha@umich.edu

\*\*Harvard Business School and Digital Reskilling Lab; jtamayo@hbs.edu

<sup>††</sup>Harvard Kennedy School and NBER. david\_deming@harvard.edu

<sup>‡‡</sup>Corresponding author Ben Weidmann (benweidmann@hks.harvard.edu). Bhalotra acknowledges funding from ESRC Centre for Microsocial Change (ESRC Award ES/S012486/1) at the University of Essex and the ESRC CAGE Centre at Warwick [grant number ES/Z504701/1]. This was supplemented by funds awarded to Deming and Weidmann by Schmidt Futures and to Vecchi 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 (HSS-REC 163/21-22). The Pre-Analysis Plan was submitted to the AEA registry prior to data collection (AEARCTR-0012258). We are grateful to Nathan Cannen, Vahid Moghani and Jordi Blanes i Vidal for helpful comments.

# 1 Introduction

Management matters greatly for economic performance (Bloom and Van Reenen, 2007; Bloom et al., 2016; Bruhn et al., 2018; Gosnell et al., 2020; Giorcelli, 2019; Metcalfe et al., 2023b). There are large and persistent productivity differences between managers within firms and across countries (Lazear et al., 2015; Adhvaryu et al., 2023; Bender et al., 2018; Metcalfe et al., 2023a). Good managers increase productivity through many channels, including monitoring performance, training and motivating staff, and reallocating workers to better roles or job tasks (Metcalfe et al., 2023a; Minni, 2023; Fenizia, 2022; Adhvaryu et al., 2023; Benson and Shaw, 2025; Dessein et al., 2024).

How do 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 and cognitive ability. Research has shown that these characteristics exhibit positive associations with manager performance in the field (Javalagi et al., 2024; Judge et al., 2002, 2004).

However, existing research on manager selection suffers from two main issues. First, managers are not randomly assigned to teams, which makes it difficult to causally identify managerial performance.<sup>1</sup> Second, managers in the field are a highly nonrandom sample of the population. This may lead to incorrect inferences about the characteristics that truly improve performance. For example, Benson et al. (2019) find that sales managers are selected based on their performance as line workers, i.e. according to the “Peter Principle”.<sup>2</sup> Promotions based on the Peter Principle can induce a negative correlation between a worker’s performance and their subsequent performance as a manager, even though the causal impact of increasing worker skills among managers is likely positive (Benson et al., 2019).

This paper introduces a novel method for identifying effective managers. We use a large pre-registered lab experiment (n=555) to make progress on the two measurement limitations mentioned above. First, we causally identify managerial contributions by randomly assigning each manager to multiple teams and controlling for task-specific performance of managers and workers, which isolates management skill from productive skill.<sup>3</sup> We design a purpose-built collaborative task that emulates real-world team

---

<sup>1</sup>A relatively small number of field studies use manager rotation for identification (Lazear et al., 2015; Minni, 2023; Metcalfe et al., 2023a; Fenizia, 2022; Wells, 2020; Giardili et al., 2022). We follow this quasi-experimental approach in our field study of retail managers (see section 4). The assumptions required to identify the causal effect of managers in these settings are demanding, limiting the scope of this approach.

<sup>2</sup>The Peter Principle states that employees are promoted to their level of incompetence (Peter and Hull, 1969).

<sup>3</sup>A related approach is used by Weidmann and Deming (2021) to identify workers who are good team

production by requiring managers to coordinate, monitor and motivate workers. Over the course of the experiment managers are randomly assigned to four different 3-person teams. Intuitively, ‘good managers’ consistently cause their teams to produce more than the sum of their parts. We find that the impact of good management is relatively large: a one standard deviation (SD) increase in managerial skill improves team performance by 0.22 SD conditional upon the productive skills of all workers including the manager. To put these figures into context, we find that *overall* manager quality matters about as much to team performance as the combined productive capacity of workers.

Second, we investigate whether people who self-select into management roles perform better on the job than people who reveal a weaker preference for managerial roles. After describing the tasks and compensation structure of the manager and worker roles, we elicit the preference of participants to be a manager on a 1-10 scale. Half of all groups were then randomly assigned to a “self-promotion” treatment where participants with the strongest preferences to be a manager were given the role. The other half of managers were drawn randomly from the participant pool.<sup>4</sup> Using this experimental variation, we find that teams with self-promoted managers perform 0.1 SD less well than teams with randomly assigned managers. This magnitude is roughly equivalent to being assigned a manager with a one standard deviation lower fluid IQ score.

We find that people who express a preference for managerial roles - who we call ‘self-promoters’ - overestimate their social skills relative to an objective test of emotional perceptiveness called the Reading the Mind in the Eyes Test (RMET).<sup>5</sup>

Among self-promoted managers, we find a negative relationship between self-reported people skills and managerial performance. In contrast, randomly selected managers do not overestimate their social skills, and we find no negative relationship between self-reported people skills and managerial performance.

What characteristics *do* predict managerial success in the lab? To answer this question we focus on the sample of “lottery managers” as the associations between manager performance and manager characteristics in the “self-promoted” arm are moderated by the filter of self-selection. We identify two consistent predictors of managerial contributions: economic decision-making skill and fluid intelligence.<sup>6</sup> These relationships are

---

players. We build on this method by introducing hierarchical teams and developing a novel collaborative teamwork task and evaluating our method using novel field data.

<sup>4</sup>Each participant was only given one role throughout the experiment.

<sup>5</sup>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; Huffman et al., 2022).

<sup>6</sup>Economic decision-making skill is defined and operationalized in Caplin et al. (2024) and focuses on the ability to strategically allocate scarce attention in a resource allocation task. Fluid intelligence is measured using Ravens Advanced Progressive Matrices.

robust to a wide range of controls, including age, gender, education, work experience, emotional perceptiveness, and personality traits including risk preferences. No other characteristics strongly predict managerial performance.

We then conduct two separate field studies to validate our experimental results and evaluate their economic importance. First, we use publicly available LinkedIn data on our lab-experiment participants to test whether managerial performance measured in the lab corresponds to meaningful differences in actual career progression. While our sample is limited to early-career workers (mean age = 27.5; mean work experience = 4.9 years) we find that a 1 SD increase in the lab measure of managerial performance is associated with 0.16 more real-world promotions per year ( $p < 0.001$ ,  $n = 73$ ; base rate of = 0.25 promotions per year). The association is robust to controls for fluid intelligence, age, gender, ethnicity and personality. This suggests that our experimental measure of managerial skill captures something predictive of career progression above and beyond traditional predictors such as education, intelligence, or demographic factors.

Second, we examine the role of managerial skills in a firm setting, where managers have repeated, long-term interactions with staff. We recruit a sample of grocery store managers employed by a multi-billion dollar retail firm in South America. Each store manager completed a set of three skill assessments that matched the broad skills assessed *before* the lab experiment: economic decision-making skill, fluid intelligence and emotional perceptiveness. We link these results to detailed data on store performance and sales over a 28 month period.

The field setting is analogous to our lab experiment in several ways. Importantly, the middle managers we study have similar core responsibilities to managers in our lab setting: they delegate tasks to frontline workers, motivate staff, monitor performance and address potential bottlenecks (e.g. inventory stockouts). As the stores share standardized practices and procedures, manager mobility across stores represents a non-random analogue of our repeated randomization procedure. Our field setting thus provides a non-experimental measure of managerial performance that can be used to quantify the value of middle managers *and* to examine whether the broad skills that predict manager contributions in the lab also predict performance in the field.

Using an event-study design we estimate that a good manager (defined as being one SD better in terms of their managerial fixed effect) improves annual per-store sales by 4.1 million USD (26%). This improvement is broadly in line with empirical estimates from other industries, which have linked a 1 SD improvement in management quality to performance improvements of 10-22%. (Giardili et al., 2022; Fenizia, 2022; Wells, 2020).

<sup>7</sup> Notably, we find that the strongest predictor of managerial performance is economic

---

<sup>7</sup>Existing studies often report different metrics, such as the impact of moving from the 25th to the 75th

decision-making skill, which aligns with our experimental findings. A 1 SD increase in economic decision-making is associated with a 0.19 standard deviation improvement in the performance of retail store managers ( $p=0.01$ ).

We then quantify the impact of different management selection regimes. Starting with evidence from the lab, results from the random assignment arm of the experiment suggest that selecting managers based on economic decision-making skill improves performance by 0.7 standard deviations relative to self-promotion. Selecting managers based on economic decision-making skill also yields significantly better manager performance than a random assignment benchmark or selecting the most productive workers to be managers (e.g. the “Peter principle”). An analogous exercise using our field data suggests that if new managers were hired based on their economic decision-making skills, annual per-store sales would increase by 2.6 million USD (16%).

Last, we investigate what good managers do to help their teams succeed. The controlled environment of the lab offers the best explicit test of mechanisms. We find that good managers: i) monitor their workers to make sure bottlenecks are being addressed; ii) match workers to the tasks that best fit their skills; and iii) keep workers motivated. Quantitatively, monitoring and motivation are more important than allocation, although all three are substantively important. In the experiment, good managers – defined as those scoring 1 SD above average in terms of their managerial skill – make monitoring errors about half as often as other managers. Similarly, in our field study we demonstrate that good managers - those with strong economic decision-making skills - are more likely to successfully monitor stock and proactively avoid stockouts.

Our paper makes three main contributions. First, we develop a skills-based approach to assessing managerial talent that can *prospectively* identify people with strong managerial potential. Our method complements the measures commonly used in economics and psychology to study managers. In economics, one stream of research has estimated manager performance by examining variation in output when teams (or firms) are exposed to different managers (Lazear et al., 2015; Metcalfe et al., 2023a; Bertrand and Schoar, 2003; Wells, 2020; Giardili et al., 2022; Fenizia, 2022; Janke et al., 2019; Minni, 2023). While these studies have many strengths they cannot be used prospectively to identify strong managers, and have very demanding data requirements.<sup>8</sup> A second stream of research has used interviews to capture differences in management *practices* across firms (e.g., Bloom and Van Reenen, 2010; Bloom et al., 2016; Bloom and Van Reenen,

---

percentile in management quality; to make these comparable, we assume that managerial quality is normally distributed.

<sup>8</sup>These approaches also suffer from the two main measurement issues noted above: non-random assignment of managers to teams and the limitation of only studying existing managers, who are a highly non-random sample. Our experiment relaxes both these constraints.

2007). We complement this literature by showing how to identify the skill of individual managers, as distinct from firm practices.

In industrial organization psychology, a related literature examines individual differences in managerial quality, relying on three main approaches to measurement: i) self- or peer-reported managerial traits (such as personality or confidence, see (e.g. Judge et al., 2002; van den Steen, 2005; Rotemberg and Saloner, 2000)); ii) self- or peer-reported managerial behaviors and leadership styles such as transformational, transactional, authentic, or identity leadership (Bass and Avolio, 2000; Beenen et al., 2021; Jensen et al., 2019; van Dick et al., 2018; Walumbwa et al., 2008); and iii) situational judgment tests in which participants are presented with hypothetical workplace scenarios and asked multiple choice questions about the best way to respond (Chan and Schmitt, 2002; Whetzel and McDaniel, 2009). Our approach complements this literature by focusing on performance-based measures of managerial skill.<sup>9</sup>

Second, we contribute to the empirical literature that quantifies the value of managers (e.g., Metcalfe et al., 2023a; Bertrand and Schoar, 2003; Lazear et al., 2015) in both lab and field settings. In particular, we provide evidence about the value of ‘middle managers’, who have been relatively understudied compared to senior leaders and CEOs (Metcalfe et al., 2023a). In the field, our movers design suffers from the measurement and identification challenges noted above, for example the non-random assignment of managers to teams. We therefore compare our estimates of the importance of managers to cleanly-identified estimates about the relative importance of managers and workers in the lab. Encouragingly, the lab and field studies produce similar results: in both cases the importance of manager quality and the *overall* quality of workers are roughly equivalent.<sup>10</sup> Similarly, the extent to which variation in manager quality is explained by skill measures and demographic characteristics is very similar in the lab and the field.

Third, we extend the literature examining the characteristics and behaviours of good managers (e.g., Javalagi et al., 2024; Judge et al., 2002, 2004), and the importance of leader selection mechanisms (e.g., Berger et al., 2020; Erkal et al., 2022; Englmaier et al., 2024). We find that single best predictor of managerial performance in both the

---

<sup>9</sup>Similarly, our experimental design complements the way in which managerial quality has been studied in experimental psychology. These studies face two common limitations: first, they do not causally isolate the contribution that individual managers make to teams; second, they typically rely on tasks that are unincentivised and not designed to mimic the skill demands of real-world management positions.

<sup>10</sup>In the lab, we find that the total effect of a 1 SD improvement in manager performance is 0.28, compared to 0.26 standard deviations for a 1 SD improvement in the average productive skills of workers (a ratio of 1.1x). In the field, we estimate that the impact on sales of a 1 SD improvement in manager quality increases sales 25 percent, compared to an 18 percent increase for a 1 SD improvement in the average quality of workers across the store (a ratio of 1.4x). In related work, Lazear et al. (2015) estimate that, across manager-worker dyads, worker effects are comparable - but somewhat larger than - manager effects (by a ratio of 1.3x)

lab and the field is a theoretically grounded measure of ‘economic decision-making’ skill (Caplin et al., 2024). Managerial effectiveness is not strongly predicted by personality and demographic traits. Notably, we also find that people with strong preferences to be in charge perform marginally worse than managers appointed at random. Overall, this suggests that a policy of proactively engaging the widest possible set of candidates and screening them based on measures of skill could substantially improve managerial quality.

The paper proceeds as follows. Section 2 describes our approach to identifying good managers in the lab and Section 3 reports our experimental findings. Section 4 presents two tests of external validity and provides initial quantitative evidence about whether our experimental results are economically meaningful. Section 5 explores the mechanisms associated with good management in the lab and the field. Section 6 concludes.

## 2 Experimentally identifying good managers

### 2.1 Identification strategy

#### Notation and setup

Let individuals be indexed by  $i = 1, \dots, n$ . Individuals are randomly assigned to groups of three people, with groups indexed by  $g$ . Let  $I_{ig}$  be a binary indicator equal to one if participant  $i$  is in group  $g$  and zero otherwise. The experiment contains two roles: ”manager” and ”worker”.<sup>11</sup> 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  $g$ .<sup>12</sup> 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 tests completed by individuals before the group session);  $G_g$  denotes the performance of group  $g$  on the Collaborative Production Task.<sup>13</sup> 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  $g$ :

$$G_g = \gamma_M \sum_i X_i M_{ig} + \gamma_W \sum_i X_i W_{ig} + \epsilon_g \quad (1)$$

$$\epsilon_g \sim N(0, \sigma_G^2).$$

<sup>11</sup>In the experimental materials we referred to ”workers” as ”team members”.

<sup>12</sup>Participants are either workers *or* managers, with roles persisting throughout the experiment

<sup>13</sup>Measures are described in section 2.3 and 2.4 and detailed in appendices A.1 and A.2. 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 a mean of 0 and a standard deviation of 1.



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

### Estimating manager performance

The residuals  $\epsilon_g$  in equation (1) can be viewed as a measure of group performance that adjusts for 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  $\epsilon_g$  arises from unmeasured individual attributes such as management skill, or from idiosyncratic match effects. However, by randomly assigning managers to multiple groups we can estimate  $\alpha_i$  - defined as the average causal impact manager  $i$  has on group performance after controlling for individual differences in productive skill - by averaging the residuals:

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

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

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

We start by estimating  $\sigma_\alpha$ , the standard deviation of the  $\alpha_i$  estimates. In model (3),  $\hat{\epsilon}_{gi}$  is a vector of skill-adjusted group performance,  $\alpha_i$  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. We use an analogous framework to estimate ”worker effects” (see Appendix A.3 for details).

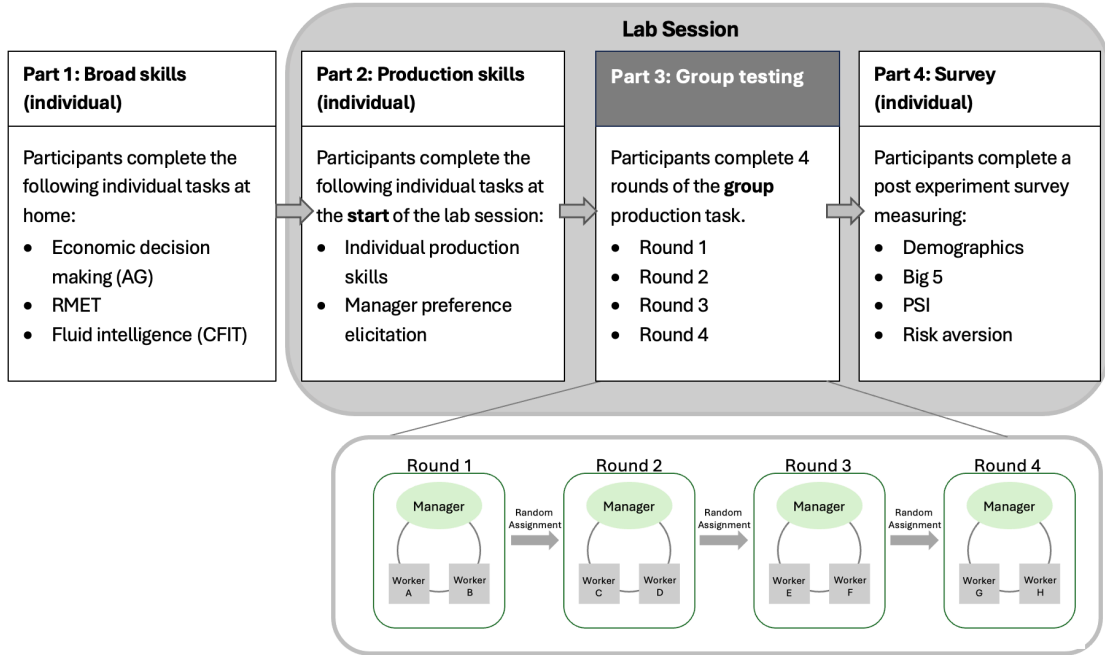
For inference, we test the null hypothesis that managers have no impact on their teams after controlling for each team’s endowment of productive skill (i.e.  $H_0 : (\sigma_\alpha) = 0$ ). Our pre-registered inferential approach is to calculate p-values using randomization inference

(see Appendix A.4 for details). For robustness, we also report alternative estimates of uncertainty using a Wald estimator and Profile Likelihood estimates.<sup>14</sup>

## 2.2 Description of experiment

Our experiment was pre-registered at the AEA RCT registry.<sup>15</sup> Figure 1 presents an overview of the experiment. Our design centers on a novel group task called the Collaborative Production Task. The task, described in section 2.3, mimics real-world team production by requiring managers to coordinate, monitor, and motivate workers. During the experiment, each manager is randomly assigned to four different teams of three people (one manager; two workers). Every time a manager is allocated to a team, they work on a parallel version of the Collaborative Production Task.

Figure 1: Overview of Experiment



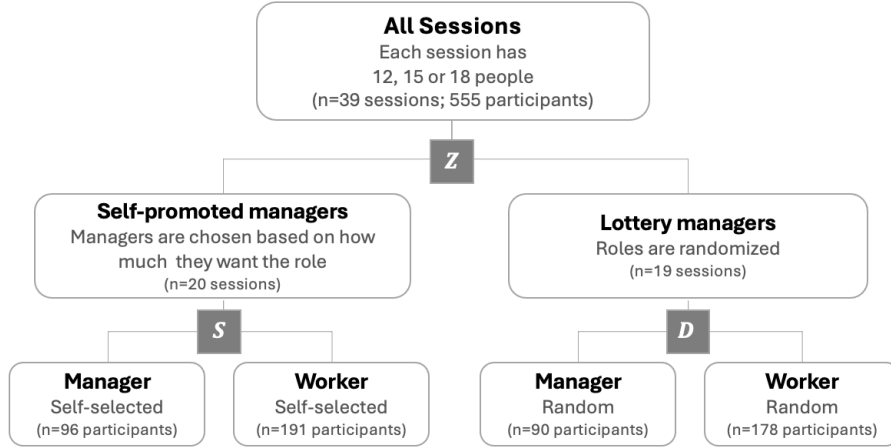
**Notes:** the figure describes the experiment from the perspective of participants. In the group testing phase each ‘round’ involves working in one group of three people. Each round involves a parallel version of the Collaborative Production Task, described in Section 2.3. The RMET is the Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001). Economic decision-making is measured by the Assignment Game (Caplin et al., 2024). Ravens is Ravens Advanced Progressive Matrices. PSI is the Political Skill Inventory, and Risk Aversion is measured by a single question asking about risk preferences.

<sup>14</sup>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).

<sup>15</sup>AEARCTR-0012258. Any deviations are flagged in footnotes.

The experiment has two levels of randomization. We randomly assign participants to teams (repeatedly) *and* we randomly vary the way in which managers are selected. Half the managers are chosen via a lottery, while the other half are self-promoted based on their stated willingness to be in charge. The manager selection mechanism is randomized at the level of ‘lab session’: 19 sessions had managers selected by lottery and 20 sessions had self-promoted managers. The randomization scheme is summarized in Figure 2.

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 selected.  $S$  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.

The experiment had two main phases: initial individual testing followed by group testing. The individual phase measured potential predictors of manager skill, including fluid intelligence (Ravens), emotional perceptiveness (RMET, Baron-Cohen et al. 2001), economic decision-making skills (Caplin et al., 2024), and participants’ preference for being a manager. We also assessed participants’ ‘productive capacity’ using individual tests analogous to components of the group task.<sup>16</sup> Group assessments were conducted at the Essex lab in the UK. Sessions involved 12, 15, or 18 participants and lasted around two hours in total.<sup>17</sup> Sessions were conducted in the morning or afternoon and every session was randomly assigned to one of two treatments governing the way in which managers were selected (‘lottery’; ‘self-promotion’).

<sup>16</sup>Individual skill tests were conducted *before* the group testing. In addition, we conducted a post-experiment survey to measure individual differences in personality and demographic characteristics.

<sup>17</sup>To minimize the chances that the same people worked together multiple times we followed a randomization scheme that minimized repeat interactions. In sessions with 15 and 18 participants, the scheme ensured that managers and workers never worked with the same worker more than once. In sessions with 12 participants, our randomization scheme ensured that team members would work together a maximum of two times.

Before group testing began, all participants rated their preference for being a manager on a scale of 1-10. They were given with detailed information about the role of the manager (see section 2.6 for details). Participants were informed that managers would be responsible for directing the group, communicating with team members, delegating, and motivating. 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. Participants did not know which treatment they were in when their managerial preference were elicited.

## 2.3 Group task

The group task was designed to meet multiple criteria, including having a distinct role for managers which replicated real-world managerial demands (for a full discussion of design considerations see Appendix A.2). In the task, groups are required to work on three question modules: numerical, spatial, and analytical reasoning. 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 answer. Each person in the team, including the manager, 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., Hirschleifer 1983).<sup>18</sup>

Each group of three sat next to each other in the lab, with the manager in the middle and a worker on either side. Before the task began, participants were informed that they were allowed to talk to each other, and teams communicated freely throughout the task. To avoid cross-team communication or interference, each team was 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 introducing themselves, groups had time to strategize about how they wanted to tackle the task. The timer began after managers had entered their initial

---

<sup>18</sup>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. Similarly, in manufacturing, a single absent or broken component can halt the entire production line. A manager is generally required to coordinate across teams to ensure the successful production of the combined product. The weakest-link setup is relevant for practical purposes because such coordination is common in economic production (Riedl et al., 2015).

task assignments. At this point 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.

### **The role of the manager**

We designed the task so that the “manager” role was both clearly differentiated from “workers”, and essential for group success (Brass, 1984). Managers had multiple 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) was allocated to a module. Managers were able to allocate more than one person to any given module.<sup>19</sup> If, for instance, a manager allocated all three participants to the ‘numerical’ module, each participant would work on a different numerical 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 had the ability to review this information about the skill profile of their teammates throughout the task. Managers also had constant access to a table describing the current allocation of tasks among their team.

Second, *managers monitored progress* throughout the task. This was especially important given the ‘weakest-link’ 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 had the lowest score. In contrast, workers are able to see the score of the module they are currently working on.<sup>20</sup> 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 swept up working on the module they assigned to themselves.

Last, *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 workers. These responsibilities were specifically incorporated into the task as they are common managerial duties, particularly for middle managers. Our focus on middle managers - as distinct from senior leaders and CEOs - was deliberate, as this group is relatively neglected in the literature (Hoffman and

---

<sup>19</sup>With three people and three modules, this allowed for 27 possible allocations.

<sup>20</sup>If two workers are working on the same module, they will both see that module’s score (rather than their individual contribution to the module).

Stanton, 2024; Metcalfe et al., 2023a).<sup>21</sup>

## 2.4 Individual tasks

### 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.<sup>22</sup> This provided us with a set of measures we could use to assess the productive capacity of any randomly assembled group.<sup>23</sup>

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.

### Broad measures of individual skill

We measured three broad skills that might predict managerial performance: economic decision-making; fluid intelligence; and emotional perceptiveness. Importantly, we asked participants to complete these assessments well before they came to the lab for group testing.<sup>24</sup> We measured economic decision-making skill - defined by Caplin et al. (2024) as the ability to make good resource allocation decisions - using the Assignment Game (Caplin et al., 2024). The game requires participants to deal with an attentionally-demanding numerical environment, understand comparative advantage intuitively, and avoid biases that undermine numerical decision-making (e.g. anchoring). Fluid intelligence was measured with a set of Ravens Advanced Progressive Matrices, a widely used assessment of the ability to solve novel problems. We also measured social skill using the Reading the Mind in the Eyes Test, or 'RMET' (Baron-Cohen et al., 2001). This psychometrically validated test of emotional perceptiveness presents participants with photos of faces, cropped so that only the eyes are visible. For each set of eyes,

---

<sup>21</sup>Our behaviorally complex task also extends the existing lab-based literature on managerial/leader impact (e.g., Brandts and Cooper, 2007; Brandts et al., 2015; Potters et al., 2007; Güth et al., 2007) in which the leader/manager role typically involves acting as the first mover in public goods or coordination games.

<sup>22</sup>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 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.

<sup>23</sup>Example problems from the individual tests of numerical, spatial and analytical reasoning are in Appendix A.1.

<sup>24</sup>This mitigates the concern that the correlation between these tests and group performance is driven by participants' mood or energy levels on the day of the group task.

participants are asked to choose which emotion, from four options, best describes the emotion being expressed.<sup>25</sup>

### **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 Political Skill Inventory (PSI) described in Ferris et al. (2005).<sup>26</sup> 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).<sup>27</sup>

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 (for example Dohmen et al. (2011) shows that the measure is strongly predictive of actual risky behaviour).

## **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 lab sample comprises 555 individuals, forming a total of 728 groups of three across the four rounds.<sup>28</sup> 46% of the sample was female, with an average age of 25. The sample was ethnically diverse with a majority of participants identifying as Asian (including South Asian) or Asian British. 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.

Participants who completed the study were paid £35 on average, with a minimum

---

<sup>25</sup>Example items from all three tests of broad individual skills are available in Appendix A.1

<sup>26</sup>Following our pre-analysis plan, items that related to the “networking” subscale are removed, as these are not relevant to our lab setting.

<sup>27</sup>We deviate from our pre-analysis plan by excluding the Indecisiveness Inventory (II) as a control variable, due to concerns about the quality and reliability of responses to the lengthy II questionnaire.

<sup>28</sup>We don’t have data for 12 groups due to data errors, primarily stemming from one session with wifi connectivity issues. To improve precision, we conducted 39 sessions instead of the 30 originally planned in the pre-analysis plan. As noted in our pre-analysis plan, power calculations were very difficult as we were using a newly-created assessment. Results are very similar if we use only the first 30 sessions, though less precise.

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 group experiment, plus a team bonus of £4 - £12 that depended on their team’s performance in one randomly selected round.<sup>29</sup> 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 Who wants to be a manager?

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, i.e., someone who would be ‘responsible for delegating, coordinating, and making decisions’. We 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.<sup>30</sup>

Figure 3 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 right panel shows the distribution of preferences among managers in the self-promotion arm. By design, managers in this arm strongly prefer to be in charge, with almost half the managers responding with a 10 on the 1-10 scale.

---

<sup>29</sup>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.

<sup>30</sup>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,  $p=0.04$ ).



Table 1: Sample and balance

|                                       | Overall<br>sample<br>(1) | Self-promoted<br>arm<br>(2) | Lottery<br>arm<br>(3) | p-value<br>(4) |
|---------------------------------------|--------------------------|-----------------------------|-----------------------|----------------|
| <b>Demographics</b>                   |                          |                             |                       |                |
| Female (%)                            | 46.3%                    | 49.1%                       | 43.3%                 | 0.17           |
| Age mean (yrs)                        | 25.0                     | 25.1                        | 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 <sup>1</sup>    | 12.3%                    | 12.6%                       | 12.0%                 | 0.10           |
| Graduate students (%)                 | 67.6%                    | 70.0%                       | 65.0%                 | 0.22           |
| <b>Skill assessments</b>              |                          |                             |                       |                |
| Task skills                           | 0.00                     | -0.01                       | 0.01                  | 0.86           |
| Fluid intelligence (Ravens)           | 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           |
| <b>Personality and working styles</b> |                          |                             |                       |                |
| 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) <sup>2</sup>       | 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 <sup>3</sup>            | 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;<sup>1</sup>Other ethnic identity includes ‘Mixed or multiple ethnic groups’, ‘Other ethnic group’, and people who preferred not to say. <sup>2</sup>Neuroticism (Big5) is reverse coded. P-values come from t-tests comparing means in the ‘lottery’ and ‘self-promoted’ arms. <sup>3</sup>Risk appetite refers to the willingness to take risks (see section 2.4 for details).

Figure 3: Histogram of preference to be manager, by treatment arm



**Notes:** The plot represents counts ( $n = 96$  self-promoted managers;  $n = 90$  lottery managers) of 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'.

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 most strongly correlated with wanting to be in charge are extraversion, risk appetite, and being male.<sup>31</sup> 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.

<sup>31</sup>Similar to several previous studies, we also 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; Haegele 2024).

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

| Dependent variable:<br>preference to be in charge (scale of 1-10) | Regression Coefficients |                   |                  |
|---|-------------------------|-------------------|------------------|
|   | Full Sample<br>(1)      | Men<br>(2)        | Women<br>(3)     |
| <b>Demographic Characteristics</b>                                |                         |                   |                  |
| Female  | -0.46*<br>(0.27)        |                   |                  |
| Age (yrs)   | 0.00<br>(0.05)          | 0.01<br>(0.07)    | -0.04<br>(0.08)  |
| Graduate student  | 0.26<br>(0.36)          | 0.61<br>(0.51)    | -0.02<br>(0.52)  |
| Years of work experience  | 0.02<br>(0.06)          | 0.08<br>(0.08)    | -0.04<br>(0.08)  |
| <b>Skill Measures</b>   |                         |                   |                  |
| Production skills   | 0.27*<br>(0.15)         | 0.43**<br>(0.19)  | 0.03<br>(0.02)   |
| Economic decision making (AG)                                     | 0.21<br>(0.15)          | 0.60***<br>(0.21) | -0.18<br>(0.22)  |
| Fluid IQ (Ravens)   | 0.23<br>(0.15)          | 0.04<br>(0.21)    | 0.42**<br>(0.22) |
| Emotional perceptiveness (RMET)                                   | -0.13<br>(0.15)         | -0.02<br>(0.21)   | -0.17<br>(0.22)  |
| <b>Personality and Working Styles</b>                             |                         |                   |                  |
| Extraversion (Big5)   | 0.49***<br>(0.15)       | 0.78***<br>(0.23) | 0.22<br>(0.20)   |
| Openness (Big5)   | 0.23<br>(0.16)          | 0.46*<br>(0.25)   | 0.15<br>(0.22)   |
| Agreeableness (Big5)  | -0.24<br>(0.17)         | -0.50**<br>(0.25) | -0.01<br>(0.24)  |
| Emotional stability (Big5)  | 0.33**<br>(0.15)        | 0.26<br>(0.23)    | 0.34<br>(0.21)   |
| Conscientiousness (Big5)  | -0.01<br>(0.16)         | 0.06<br>(0.22)    | 0.05<br>(0.24)   |
| Political Skill Inventory   | 0.09<br>(0.16)          | 0.06<br>(0.23)    | 0.20<br>(0.24)   |
| Risk appetite   | 0.42***<br>(0.15)       | 0.43**<br>(0.21)  | 0.39*<br>(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 ‘willingness to be a manager’, participants were informed that they should choose the role that ‘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 1 if participants were graduate students, and 0 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 (2022). Significance levels are denoted by: \*\*\*  $p < 0.01$ ; \*\*  $p < 0.05$ ; \*  $p < 0.1$

### 3 Experimental Evidence

This section presents our main pre-registered results, deviations are indicated in text of footnotes.

#### 3.1 Do managers matter?

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 multiple inference methods. 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' (i.e. the impact of receiving a manager who is 1SD above average) is 0.22 standard deviations ( $p = 0.03$ ). 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. To illustrate the *total* average causal contribution individual managers 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 0.22 SD to 0.28 SD.<sup>32</sup> Comparing the results in columns 1 and 2 also allows us to contextualize the importance of overall manager quality with the value of having more productive workers. We find that the overall impact of a 1 SD better manager is 0.28 SD (column 2), while increasing the productive skills of the workers by 1 SD increases team production by 0.27 SD (column 1). This suggests that manager quality matters about as much to team production as the *overall* productive capacity of workers.

---

<sup>32</sup>Removing the condition also dramatically increases the magnitude of the worker effect, from 0.04 SD (ns) to 0.21 SD ( $p = 0.04$ ). 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.

Table 3: Estimating the magnitude of manager and worker effects

|  | Dependent variable: Group Performance (G) |           |           |           |           |           |           |           |           |
|--|---|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
|  | (1)                                       | (2)       | (3)       | (4)       | (5)       | (6)       | (7)       | (8)       | (9)       |
| <b>Manager effect (<math>\hat{\sigma}_\alpha</math>)</b> | 0.221                                     | 0.284     | 0.222     | 0.218     | 0.219     | 0.214     | 0.215     | 0.179     | 0.253     |
| [Randomization inference]                                | [0.03]                                    | [<0.01]   | [0.04]    | [0.05]    | [0.04]    | [0.05]    | [0.04]    | [0.194]   | [0.05]    |
| {Wald}   | {< 0.005}                                 | {< 0.005} | {< 0.005} | {< 0.005} | {< 0.005} | {< 0.005} | {< 0.005} | {< 0.005} | {< 0.005} |
| (Profile likelihood)                                     | (0.05)                                    | (0.01)    | (0.05)    | (0.05)    | (0.05)    | (0.05)    | (0.05)    | (0.21)    | (0.06)    |
| <b>Worker effect (<math>\hat{\sigma}_\Omega</math>)</b>  | 0.036                                     | 0.205     | 0         | 0         | 0         | 0.074     | 0.057     | 0         | 0         |
| [Randomization inference]                                | [0.49]                                    | [0.04]    | [-]       | [-]       | [-]       | [0.48]    | [0.48]    | [-]       | [-]       |
| {Wald}   | {0.40}                                    | {< 0.005} | {-}       | {-}       | {-}       | {0.02}    | {0.40}    | {-}       | {-}       |
| (Profile likelihood)                                     | (0.49)                                    | (0.03)    | (-)       | (-)       | (-)       | (0.40)    | (0.49)    | (-)       | (-)       |
| <b>Controls</b>  |   |           |           |           |           |           |           |           |           |
| Manager's production skills <sup>1</sup>                 | 0.214                                     |           | 0.206     | *         | *         | 0.220     | 0.216     |           |           |
|  | (0.040)                                   |           | (0.041)   | *         | *         | (0.040)   | (0.040)   |           |           |
| Workers' production skills <sup>2</sup>                  | 0.265                                     |           | 0.242     | *         | *         | 0.266     | 0.265     |           |           |
|  | (0.035)                                   |           | (0.040)   | *         | *         | (0.035)   | (0.035)   |           |           |
| Variance of team production skills <sup>3</sup>          |   |           | x         |           |           |           |           |           |           |
| Granular production skills (both) <sup>4</sup>           |   |           |           | x         |           |           |           |           |           |
| Non-linear production skills (both) <sup>5</sup>         |   |           |           |           | x         |           |           |           |           |
| Manager familiarity with other participants <sup>6</sup> |   |           |           |           |           | x         |           |           |           |
| Manager risk appetite <sup>7</sup>                       |   |           |           |           |           |           | x         |           |           |
| <b>Counts</b>  |   |           |           |           |           |           |           |           |           |
| Groups [4 rounds per person]                             | 728                                       | 728       | 700       | 700       | 700       | 700       | 700       | 360       | 360       |
| Managers   | 186                                       | 186       | 176       | 176       | 176       | 176       | 176       | 90        | 90        |
| Workers  | 369                                       | 369       | 357       | 357       | 357       | 357       | 357       | 179       | 179       |
| Manager Sample   | All                                       | All       | All       | All       | All       | All       | All       | Lottery   | Lottery   |

*Notes:* The dependent variable is group scores,  $G_g$ . 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, the null hypothesis being tested in all cases that  $\hat{\sigma}_x = 0$ . <sup>1</sup>For details on our inference procedure, see Appendix A.4. Manager production skills are defined at the individual level as the standardized score across the numerical, analytical and spatial tasks (SD=1, mean=0). <sup>2</sup>Worker production skills are defined at the group level as the mean score across workers and modules (and standardized so that this variable has SD=1 and mean=0 across all groups). <sup>3</sup>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). <sup>4</sup>Granular production skills are the scores on the numerical, analytical and spatial tasks for both managers and workers. 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. Lastly, note that the estimate of 0 for the worker effect in column (6) comes from multilevel model (3), which estimates zero variance at the level of individual workers. <sup>5</sup> controls for non-linear individual skills of both managers and workers; <sup>6</sup>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. <sup>7</sup>Manager risk appetite is measured on a scale of 1-10.

Columns 3 to 9 present robustness checks, demonstrating that our results are unlikely to be driven by specifics of the conditioning model presented in equation (1). We add controls for: a measure of the variance of productivity within the team (col 3); a full set of fine-grained measures capturing each team member’s individual skills (col 4); non-linear individual skills of both managers and workers (col 5); whether managers knew any of the teammates (col 6); manager risk appetite (col 7); and restricting the sample to the lottery managers only (col 8-9).<sup>33</sup> Estimated manager effects are robust to all these controls.

Following our pre-analysis plan, we also test robustness using a leave-one-out procedure in which we hold out data from one of the four experimental rounds, then calculate the average causal contributions of individual managers ( $\alpha_i^{LOO}$ ) and workers ( $\Omega_i^{LOO}$ ) using the remaining three rounds. We then assess whether the estimated manager and worker effects predict whether a group will be successful in the holdout round. We repeat this for each round (see Appendix B.1 for details). We find that manager contributions from 3 rounds of data predict team performance in the holdout round ( $p < 0.01$ ). 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.

### 3.2 Self-promoters perform worse than lottery managers

Next, we examine the performance of participants who express a strong desire to be a manager. Because the lab setting allows us to randomize 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 4 contrasts the performance of groups managed by a ‘lottery manager’ with those of a ‘self-promoted manager’. We regress group performance  $G_g$  on a binary indicator being in the ‘self-promotion’ arm. Column 1 shows that teams led by self-promoted managers do *worse* on average than teams led by lottery managers. The estimate is noisy (0.13 SD,  $p = 0.09$ ), but we can reject that self-promoters do better. 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 SD. This magnitude is meaningful: on average,

---

<sup>33</sup>Column’s 4, 8 and 9 were added as a response to referee feedback and were not pre-registered. We are grateful for the feedback.

groups with self-promoted managers perform about as poorly as groups with mean fluid intelligence almost 1 SD below average.

Why might people who prefer to be managers fail to outperform people who were randomly chosen, some of whom did not want the job? We hypothesize that self-promoted managers are overconfident, especially about their social skills. This may lead them to focus too much on their own behaviour at the expense of focusing on the skills and motivation of their teammates. We test this idea with exploratory analyses in which we first assess overconfidence on the group task. After the final round of group tasks, we ask managers to reflect on whether they had performed “much worse,” “worse,” “average,” “better,” or “much better” than their peers. We computed an individual-level measure of overconfidence by regressing self-reported performance on each manager’s causal contribution to the team ( $\alpha$  and capturing the residual. People who want to be in charge were much more overconfident than 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); Huffman et al. (2022)). Although gender does not moderate the relationship between wanting to be in charge and overconfidence, we find that men are significantly more overconfident on average than women (the mean difference is 0.35 SD,  $p=0.02$ ), in line with evidence from Exley and Kessler (2022); Tradenta et al. (2025) who similarly find gender differences in self promotion. Random selection of managers will thus tend to result in inclusion of women, also see Goodall and Osterloh (2015).

Second, we find that self-promoted managers are particularly overconfident about their social skills. Among self-promoters we find a strong negative correlation between self-reported people skills and performance on the RMET (correlation =  $-0.37$ ,  $p < 0.001$ ).<sup>34</sup> However, the relationship was not significant for lottery managers. Last, we note that managers who were overconfident about their task performance scored worse on the RMET (correlation =  $-0.33$ ,  $p < 0.001$ ). Overall, we find suggestive evidence that self-promoted managers are overconfident about their performance in general, and about their social skills in particular.

---

<sup>34</sup>Self-reported people skills are calculated as the average of extraversion and Political Skill. Heck et al. (2024) find a negative relationship between self-reported and skill-based tests of social intelligence.

Table 4: 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)                             | (9)                             |
|---|----------------------------------|----------------------------------|----------------------------------|----------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| <b>Self-promoted vs lottery<br/>(standard errors)</b> | <b>-0.128*</b><br><b>(0.074)</b> | <b>-0.127*</b><br><b>(0.070)</b> | <b>-0.119*</b><br><b>(0.069)</b> | <b>-0.115*</b><br><b>(0.069)</b> | <b>-0.105</b><br><b>(0.069)</b> | <b>-0.100</b><br><b>(0.070)</b> | <b>-0.096</b><br><b>(0.070)</b> | <b>-0.098</b><br><b>(0.071)</b> | <b>-0.113</b><br><b>(0.074)</b> |
| Production skills <sup>1</sup>                        |                                  | 0.189                            | 0.146                            | 0.142                            | 0.138                           | 0.138                           | 0.138                           | 0.140                           | 0.134                           |
| Fluid IQ (Ravens) <sup>2</sup>                        |                                  |                                  | 0.104                            | 0.098                            | 0.094                           | 0.104                           | 0.104                           | 0.103                           | 0.101                           |
| Economic Decision-Making (AG) <sup>2</sup>            |                                  |                                  |                                  | 0.019                            | 0.011                           | 0.010                           | 0.010                           | 0.009                           | 0.008                           |
| Emotional Perceptiveness (RMET) <sup>2</sup>          |                                  |                                  |                                  |                                  | 0.028                           | 0.016                           | 0.019                           | 0.023                           | 0.018                           |
| Risk appetite <sup>3</sup>                            |                                  |                                  |                                  |                                  |                                 | -0.003                          | -0.001                          | 0.001                           | 0.021                           |
| Know others in experiment <sup>4</sup>                |                                  |                                  |                                  |                                  |                                 |                                 | 0.076                           | 0.073                           | 0.090                           |
| Education and work experience <sup>5</sup>            |                                  |                                  |                                  |                                  |                                 |                                 |                                 | x                               | x                               |
| Big5 and political savvy <sup>6</sup>                 |                                  |                                  |                                  |                                  |                                 |                                 |                                 |                                 | x                               |
| Obs   | 728                              | 728                              | 728                              | 728                              | 728                             | 700                             | 700                             | 700                             | 700                             |
| Adjusted R <sup>2</sup>                               | 0.004                            | 0.108                            | 0.136                            | 0.137                            | 0.139                           | 0.146                           | 0.147                           | 0.147                           | 0.158                           |

*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. <sup>1</sup>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. <sup>2</sup>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 Ravens (standardized to have mean=0 and SD=1). <sup>3</sup>Risk appetite is a self-reported measure of the risk tolerance of the group’s manager, on a scale of 1-10. <sup>4</sup>Know others in experiment is a binary indicator at the group level equal to one if the manager knew either of the workers. <sup>5</sup>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. <sup>6</sup>Big5 and political savvy are self-reported measures of personality and political skill, described in section 2.4. Significance: \*\*\* $p < 0.01$ ; \*\* $p < 0.05$ ; \* $p < 0.1$ .



### 3.3 What predicts being a good manager in the lab?

Table 5 explores the characteristics associated with being a good manager in the experiment. Table 5 separately reports predictors of management performance for lottery managers (column 1) as well as 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 (Ravens) and economic decision-making (Assignment Game), both of which are statistically significant at the less than one percent level.<sup>35</sup> These associations are robust to a wide range of controls, including age, gender, education and work experience, along with measures of emotional perceptiveness and personality traits including risk aversion (see Table B.2 in Appendix B for details).<sup>36</sup>

Among self-promoted managers we find negative correlations between management performance and both self-reported extraversion ( $\rho = -0.23$ ,  $p < 0.05$ ) and self-reported political skill ( $\rho = -0.24$ ,  $p < 0.05$ ). 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.

### 3.4 Quantifying the impact of manager selection mechanisms in the lab

We quantify the benefits of skill-based promotions by 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.<sup>37</sup>

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 one-third of participants as managers (as 1 in 3 people in the experiment are managers). To estimate the average quality of managers under the Peter Principle we first rank participants in the lottery arm according to their scores on the individual tests of productivity (assessed before the group tests) and identify the top tercile. The average manager quality among this group is an unbiased estimate of the average manager quality under a regime where managers are

<sup>35</sup>Note that these correlations are exploratory, as per our analysis plan (Parker and Weir, 2022).

<sup>36</sup>Managers make decisions under risk and uncertainty, and if cognitive or decision-making ability is correlated with risk preferences, it becomes relevant to control for risk preferences (Dohmen et al., 2018).

<sup>37</sup>This subsection was not pre-registered as it is exploratory in nature.

Table 5: Correlates of management performance

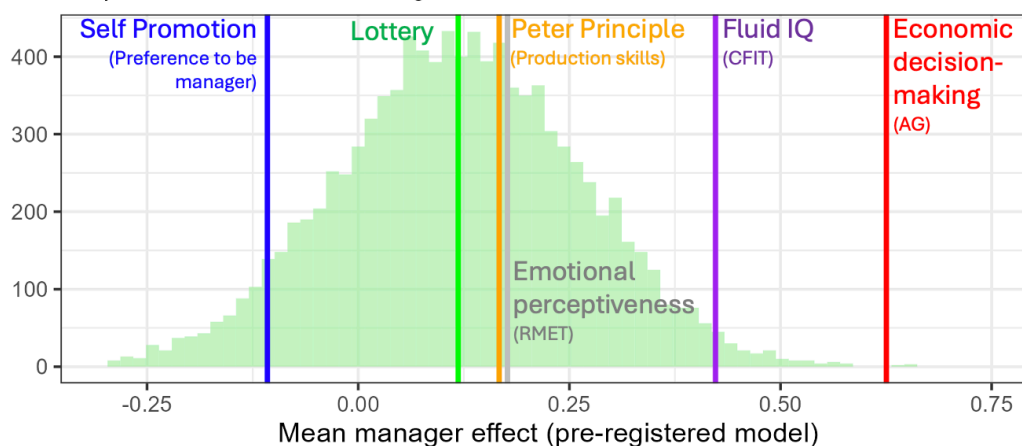
| <b>Correlation with manager contributions</b>   |  |  |  |
|---|--|--|--|
| Pairwise (marginal) correlation with $\hat{a}_i$ (our measure of manager contribution, conditional on the team's endowment of production, as per our pre-specified model <sup>1</sup> ) |  |  |  |
|   | <b>Lottery<br/>managers</b><br>n=90<br>(1) | <b>Self-promoted<br/>managers</b><br>n=96<br>(2) | <b>Full<br/>sample</b><br>n=186<br>(3) |
| <b>Skill assessments</b>  |  |  |  |
| Fluid intelligence (CFIT)   | 0.22<br>(0.04)                             | 0.29<br>(<0.01)                                  | 0.25<br>(<0.01)                        |
| Economic Decision-making (AG)   | 0.23<br>(0.03)                             | 0.10<br>(0.34)                                   | 0.16<br>(0.03)                         |
| Emotional perceptiveness (RMET)   | -0.02<br>(0.87)                            | 0.19<br>(0.07)                                   | 0.09<br>(0.21)                         |
| <b>Personality and working styles</b>   |  |  |  |
| Extraversion (Big5)   | -0.07<br>(0.51)                            | -0.23<br>(0.03)                                  | -0.14<br>(0.06)                        |
| Openness (Big5)   | -0.19<br>(0.08)                            | -0.04<br>(0.73)                                  | -0.12<br>(0.13)                        |
| Agreeableness (Big5)  | -0.18<br>(0.10)                            | 0.04<br>(0.70)                                   | -0.05<br>(0.52)                        |
| Neuroticism (Big5)  | 0.10<br>(0.37)                             | 0.03<br>(0.76)                                   | 0.05<br>(0.50)                         |
| Conscientiousness (Big5)  | 0.00<br>(0.99)                             | -0.02<br>(0.88)                                  | -0.02<br>(0.84)                        |
| Political Savvy (PSI)   | -0.04<br>(0.72)                            | -0.24<br>(0.02)                                  | -0.14<br>(0.06)                        |
| Risk appetite   | 0.07<br>(0.53)                             | -0.12<br>(0.26)                                  | -0.03<br>(0.69)                        |
| <b>Demographics</b>   |  |  |  |
| Age   | 0.00<br>(0.99)                             | 0.15<br>(0.16)                                   | 0.07<br>(0.35)                         |
| Female  | -0.10<br>(0.38)                            | -0.10<br>(0.33)                                  | -0.09<br>(0.25)                        |
| Years of work experience  | -0.07<br>(0.54)                            | 0.07<br>(0.54)                                   | 0.00<br>(0.99)                         |

*Notes:* <sup>1</sup>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, 2022).

promoted based on individual productivity.<sup>38</sup> Analogously, we can look at any other individual skill as a basis for manager selection.

Figure 4 compares the results of six selection mechanisms. We examine self-promotion, lottery, and choosing managers based on four skills: fluid IQ (Ravens), economic decision-making (AG), emotional perceptiveness (RMET), and individual production skills (i.e., the Peter Principle). 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 0.7 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.<sup>39</sup>

Figure 4: Comparing methods of appointing managers



**Notes:** To demonstrate sampling uncertainty, Figure 4 includes a sampling distribution (in green) illustrating multiple draws 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 Ravens is not statistically significant. Finally, note that the manager effects are normalized so that, across the sample of managers, the SD of manager effects = 1, and the mean = 0.

<sup>38</sup>It is important to use the lottery arm because self-selection is correlated with skills and other characteristics. These correlations undermine the ability to estimate the manager quality under the Peter Principle (or based on other skills).

<sup>39</sup>A limitation is that these simulations ignore the possibility that team members may behave differently when they are aware they are under different selection rules. We thank a reviewer for noting this.

## 4 Field Evidence

So far we have presented a novel experimental approach to identify good managers in the lab. This section provides preliminary evidence to externally validate our results. In particular, we explore three questions:

1. Does our experimental measure of management skill predict labor-market success for individuals?
2. Do predictors of success in the lab predict managerial performance in a firm?
3. What is the economic value of selecting middle managers based on skills?

We conduct two separate studies. First we examine the labor-market experiences of our lab experimental participants using publicly available data from LinkedIn. Second, we conduct a field study of a multi-billion dollar retail firm in South America. We discuss these in turn.<sup>40</sup>

### 4.1 Labor-market outcomes of experimental participants

The first external test of our method focuses on the labor-market outcomes of our experimental participants. We hypothesize that managers who succeed in the lab will be more likely to succeed in the labor market. In particular, we hypothesize that managerial skill (as measured by  $\alpha_i$ , which controls for differences in fluid intelligence and problem-solving skill) is positively associated with the rate at which people are promoted. This hypothesis builds on the idea that young workers are initially hired based on easily observable characteristics and that subsequent promotion and hiring decisions are increasingly shaped by hard-to-measure correlates of productivity (Altonji and Pierret, 2001).

#### Data and method

From our original sample of 184 managers, we find publicly available labor-market data on LinkedIn for  $n=73$  participants.<sup>41</sup> This LinkedIn sample represents people in the early stages of their career: mean age is 27.5 and participants have an average of 4.9 years of work experience. Participants in the LinkedIn sample ( $n=73$ ) are very similar to other experimental participants ( $n=109$ ) in terms of skills, gender and managerial preferences.<sup>42</sup>

---

<sup>40</sup>The analysis in this section was not part of our experimental pre-registration. We are grateful to anonymous reviewers for encouraging us to examine field outcomes, which significantly enhance the external validity of our findings.

<sup>41</sup>Of the  $n=109$  participants who have no full-time job records on LinkedIn,  $n=36$  are students who are yet to enter the labor market;  $n=27$  have a LinkedIn page but no information on full-time work and  $n=46$  did not have a LinkedIn page.

<sup>42</sup>See Table C.1 in Appendix C for details. The only significant difference between the two samples is

The primary outcome measure we observe on LinkedIn is career progression. Our data includes 168 job transitions; 62 of these were coded as promotions.<sup>43</sup> To adjust for differences in the time participants spend in the labor market we measure time elapsed since completion of undergraduate studies, minus any periods of time when participants were engaged in full-time postgraduate study.

## Results

Participants who perform well as managers in the lab are more likely to receive promotions in the labor market. A 1 SD increase in management performance  $\alpha_i$  is associated with 0.16 more promotions per year ( $n=73$  managers,  $p=0.001$ ). ‘Good managers’ in our experiment (+1 SD) are promoted every 2.3 years on average, compared to 5.3 years for other participants ( $n=73$ ,  $p=0.03$ ). Other lab-based measures of skill are also predictive of early career promotions, albeit with smaller magnitudes: a 1 SD increase in economic decision making is associated with 0.10 more promotions per year ( $p=0.03$ ) while a 1 SD increase in fluid intelligence is associated with 0.13 more promotions per year ( $p=0.01$ ).

Overall, we find that participants who demonstrate strong management skills in the lab are substantially more likely to experience early-career promotions, and that this association is robust to controls for fluid intelligence, age, gender, ethnicity and personality (see Appendix C for details).<sup>44</sup> This suggests that our experimental measure of managerial skill is predictive of career progression and that it carries information above and beyond traditional predictors such as intelligence or demographic characteristics. Moreover, as our samples focuses on early-career workers, the analysis likely understates the predictiveness of performance-based measures of managerial skill, which will arguably become more valuable and salient as workers mature (Altonji and Pierret, 2001).

In a secondary analysis we examine whether ‘self-promoted’ and ‘lottery’ managers in our experiment present themselves in observably different ways outside the lab. First, we note that the probability of being in the LinkedIn sample is very similar for self-promoted and lottery managers:  $P(\text{LinkedIn} \mid \text{self promoter})=0.43$ ,  $P(\text{LinkedIn} \mid \text{lottery}$

---

that people with data on LinkedIn are somewhat older (mean age of 27.5 compared to 23.5 for other experimental participants).

<sup>43</sup>We code transitions as a promotion if: i) the new job title clearly signals an increase in seniority (e.g. ‘data analyst’ -> ‘senior data analyst’); or ii) a participant transitioned from a low-wage country to a high-wage country in a similar role (e.g. ‘software developer’ in Lagos -> ‘software developer’ in London)

<sup>44</sup>Following recent results from Haegele (2024) suggesting the important role of gender in wanting (and receiving) promotions, we paid particular attention to the role of gender. In our sample, we found no significant gender differences in: i) the rate at which people were promoted; and ii) the relationship between ‘managerial skills’ (measured in the lab) and promotion rates. We suspect this is due to sample size, coupled with the fact that our participants are in the early years of their careers (Haegele’s primary sample has an average age of 44).

manager)=0.39, ( $p=0.66$ ,  $n=184$ ). Next, we use the data on LinkedIn to see how people describe their skills, focusing in particular on 'management' skills. For each participant, we count the number of skills listed on their LinkedIn page related to management, e.g. 'team leadership' or 'management'. We find that self-promoters list significantly more of these skills than the lottery managers (averaging 4.1 management skills compared to 2.4 for lottery managers,  $p=0.02$ ;  $n=108$ ). This association is robust to demographic controls (age and gender). This suggests that our lab-based elicitation of managerial preferences reflect meaningful differences in how participants view their own skill profiles, and how they present themselves in the labor-market.<sup>45</sup>

## 4.2 Managerial skill in a large retail firm

We now turn to our field evaluation, which studies managers in a multi-billion dollar retail firm in South America. Our dataset comes from one country, in which the firm owns 500 grocery stores. The firm has a three-layered organizational structure: central headquarters (responsible for strategic direction, financial decisions and HR policies across stores); middle managers (who are the store managers and are the focus of our study) and front-line workers. In this sense, the teams composed of middle managers and frontline workers are akin to the teams in our lab experiment described in section 2.

This setting was chosen for two main reasons. First, stores within the firm have standardized management practices. Studying the performance of managers inside such a firm allows us to control for firm-level managerial practices (of the sort captured by e.g. Bloom et al. (2016)) and focus on the impact of manager-specific skills in enabling productivity of frontline workers.

Second, this setting has structural similarities to our lab experiment. Store managers perform similar functions to "managers" in the lab: they assign frontline workers to tasks, motivate staff, monitor performance and address potential bottlenecks (e.g., inventory stockouts) – but they are typically not in charge of hiring or firing workers.<sup>46</sup> Moreover, managers in the field regularly rotate across stores within the firm. These manager moves are the non-randomized analogue of our repeated randomization procedure: in the lab, managers are randomly assigned to multiple teams; in the field, they are cycled through multiple stores by central headquarters. As such, this field setting provides a non-experimental measure of managerial performance (using a switching de-

<sup>45</sup>The number of type of skills listed on LinkedIn communicates substantial information about labor-market experiences: recent work from Dorn et al. (2024) suggests that LinkedIn skill listings explain more variation in earnings than education and experience.

<sup>46</sup>Managers make recommendations for promotions and are consulted by HR before a promotion is made.

sign, described below) that can be used to quantify the economic contribution of middle managers to store performance and to examine whether the broad skills that predict manager contributions in the lab also predict manager value in the field.

## Data

We recruited a sample of 225 managers for whom we could observe work history, including any transitions between stores. We had the managers complete a set of skill assessments that are analogous to the ‘broad measures of individual skill’ used in the lab experiment (see section 2.4). These assessments included a Spanish-language version of the Assignment Game to measure economic decision-making skills (Caplin et al., 2024), an abbreviated set of Ravens Advanced Progressive Matrices to measure fluid intelligence, and the Reading the Mind in the Eyes Test as a performance-based proxy measure of social skill Baron-Cohen et al. (2001). In addition to these skill assessments, managers completed a detailed survey capturing their demographic characteristics (e.g. age, gender and marital status, civil status) and their work experience (type of contract, whether or not they were unionized, years of work experience). Managers receive competitive local salaries, with 30-40% performance-based incentives.<sup>47</sup> Sample statistics describing the sample of middle managers are presented in Table D.1 in the Appendix.

The manager data are linked to a panel of store-level data covering the period February 2020 to May 2022. For each store we observe monthly sales and employment (full and part-time worker numbers) along with complete records about the identity of managers and their movements.<sup>48</sup> The administrative store data also contain information about store inventory, including the frequency of stockouts.

## Empirical Strategy

Our empirical analysis starts by estimating the value added by managers in terms of log monthly sales. Following Abowd et al. (1999), we leverage the mobility of managers across stores to estimate manager and store fixed effects, with the following model:<sup>49</sup>

$$\log(\text{Sales}_{smt}) = \alpha + \theta_m + \psi_{s(m,t)} + \delta_t + \gamma \text{demo}_{st} + \varepsilon_{smt} \quad (4)$$

---

<sup>47</sup>This is analogous to the lab study, where the maximum performance incentive was 40% of total pay.

<sup>48</sup>Most stores will have more than one manager at a given time, typically with a single general manager and several department or operation-specific assistant managers. In our survey, we targeted the highest level manager available for each store. As such most stores will have a single main store manager in both the survey data and store performance panel data, but in some instances more than one of the managers from our survey sample may appear in a given store for some period of time. When necessary, we average measures across these managers.

<sup>49</sup>We control for manager demographics, time fixed effects, and the demographic composition of workers for each store where the manager worked. These store-level controls include the proportion of female workers, the average age and average tenure of workers, the distribution of civil status categories, and the proportion of unionized employees.

The dependent variable is  $\log(\text{Sales}_{smt})$ , defined as the log of total sales in store  $s$  reporting to manager store  $m$ , in month store  $t$ ;  $\theta_m$  is a fixed effect for the manager,  $\psi_s$  refers to a store-fixed effect, and  $\delta_t$  is a date (monthly) fixed effect that accounts for the seasonality of sales, as well as the general changes that the firm implements based on periodic goals. We follow Card et al. (2013) and decompose the error term  $\epsilon_{smt}$  into a match-specific component  $\eta_{sm}$ ; a store root component  $\phi_{mt}$  and a transitory error  $\nu_{smt}$ .<sup>50</sup>

## Identification

The identification of manager fixed effects relies on the assumption that the assignment of managers to stores is conditionally mean-independent of past, present, and future values of  $\epsilon_{smt}$ . This assumption allows managers to be assigned to stores based on the permanent components of managerial ability  $\theta_m$  and store components  $\psi_s$ , permitting sorting on these fixed effects. However, it excludes the possibility of managers being assigned to stores based on their match-specific component  $\eta_{sm}$  or transitory shocks to store performance  $\nu_{smt}$ . While we provide evidence supporting these assumptions below, we emphasize the point made in the Introduction, that these are very demanding identification requirements that are difficult to establish in practice - highlighting the relevance of our lab experiment.

Additionally, as discussed in Abowd et al. (1999, 2002), the manager and worker fixed effects in this model are separately identified only within “connected sets” of stores, linked by managers moving across stores.<sup>51</sup>

We identify 55 connected sets (CS) in the universe of store data and estimate equation (4) within each connected set.<sup>52</sup>

## Analysis of manager moves

Managers rotate across stores fairly regularly. Over the period covered by our data panel 12.9% of the 225 managers worked in two or more stores.<sup>53</sup> Similarly, stores change

<sup>50</sup>We also estimate a ‘naïve’ fixed effects model, that is the same as equation (4) but does not include  $\text{demo}_{st}$ . We use naive estimates as a sense check, and also to help quantify the relative importance of manager- and worker-quality.

<sup>51</sup>Note that this universe of store data includes many times more than the 225 managers for whom survey data was obtained. We leverage this larger universe to maximize identification of the manager fixed effects and later match these estimates to the survey sample at the manager level for much of our analysis.

<sup>52</sup>Relatedly, if we do not observe sufficient moves of managers across each of these trajectories between stores of varying quality we may not be able to separately obtain precise and consistent estimates of manager and store fixed effects, as discussed in Abowd et al. (1999); Andrews et al. (2008). To assess this concern, in Table D.4 we implement the bias correction procedure suggested by Andrews et al. (2008) and additionally allow for heteroskedasticity by implementing the leave-out estimator proposed by Kline et al. (2020). Both exercises yield distributions of manager and store FE with very similar moments from the baseline model, indicating that limited mobility bias is not a significant concern.

<sup>53</sup>This share of movers is consistent with previous seminal papers that have leveraged the same ana-



managers relatively frequently, providing the variation required to identify managerial effects. In our data, 26.9% of all stores experience at least one manager change.

While we do not observe why individual managers move across stores - and cannot rule out the possibility that these decisions are shaped by store characteristics - firm policies indicate that many of these moves are for administrative reasons unrelated to store performance, and do not reflect deliberate matching between manager and store characteristics. Managerial assignments across stores are centrally coordinated by headquarters, with rotations primarily aimed at fostering professional development. The underlying premise is that exposure to a variety of store environments enhances managerial capabilities. As a result, the timing and identity of incoming managers are largely exogenous to store-level conditions.

To assess the suitability of this setting for estimating manager effects, we perform several tests analyzing manager mobility and its potential correlation with store performance dynamics (Card et al., 2013; Adhvaryu et al., 2024). In particular, we follow Card et al. (2013) in presenting two main exercises to test the assumptions underlying identification of manager fixed effects. First, in Figure D.1 we confirm that the moves of managers across stores appear to be conditionally mean independent of the match-specific component, by demonstrating that the gains and losses from moving managers across stores are roughly symmetric. That is, the difference in store performance from a manager moving from a low quality (e.g., first quartile in residualized average log sales) store to a high quality (e.g., fourth quartile in residualized average log sales) store should be roughly equal and opposite to the losses from a manager moving from a high quality store to a low quality store. Indeed, the patterns presented in Figure D.1 support this assumption. Note also that the slopes just before and after the moves are almost all perfectly flat, addressing any concern that these moves follow trends in performance in origin stores or coincide with other initiatives in the destination store that affect performance. We discuss additional evidence in support of this below.

Then, to better assess quantitatively how symmetric are these gains and losses from moves, we plot in Figure D.2 the gains from moving up against the losses from moving down from all possible moves across quartiles. The estimates are indeed all close to the 45-degree line reflecting symmetry.

## Results

What is the economic impact of improving manager quality in this field context? To answer this question, we conduct an event study in which the events reflect the arrival of a manager who raises the managerial quality of the store (i.e., the quality of the arriving

---

lytical framework. See, for example, Card et al. (2013); Abowd et al. (1999); Metcalfe et al. (2023a)

manager is above the average managerial quality for the store, defined by the manager's estimated fixed effect).<sup>54</sup> The results are presented in Figure 5. We find that the arrival of a higher quality manager results in a large, immediate and significant increase in sales. Interviews with the firm confirm that an immediate impact is expected, as better managers are able to make rapid changes on multiple fronts, including pricing and adjustments to avoid stockouts (discussed below). After 4 months, we estimate that an 'above average' manager raises sales by 13 percent (SE=3 percent). We estimate that the arrival of a 'good manager' (defined as a manager who is 1 SD above average) increases sales by 25 percent (SE=6 percent) which equates to an increase in annual turnover of 4.1 million USD.<sup>55</sup>

Figure 5: Impact of 'above-average' managers on *sales*: Event Study

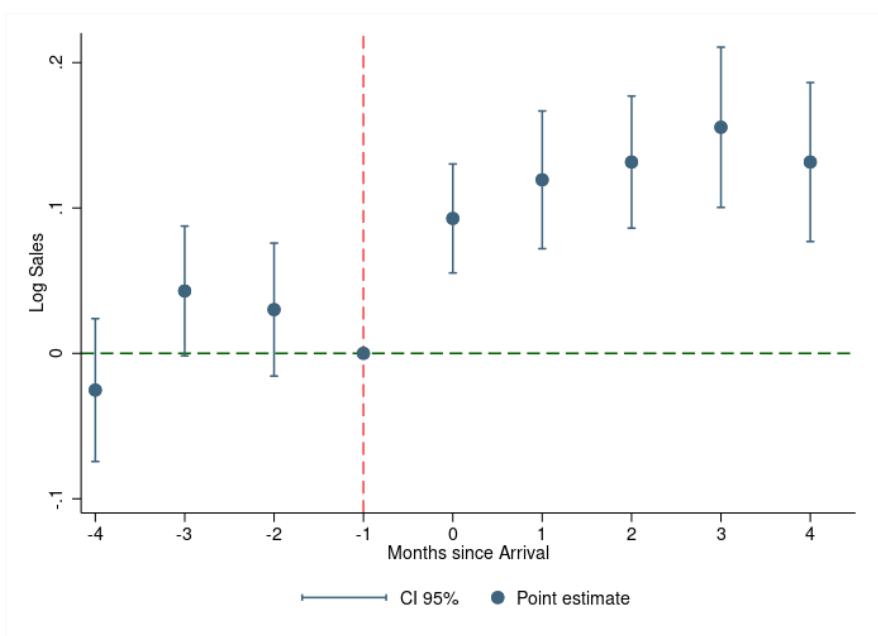


Figure 5 shows the results from estimating an event study design using the arrival of a manager who raises the managerial quality of a store. We consider an event to be when the fixed effect of the arriving managers is higher than the mean of those across pre-existing managers in the store one month before the arrival. Fixed effects are estimated according to equation 4. Standard errors are clustered at the store level. The average number of managers before the arrival is 6.1. The mean gap between average pre-existing manager and the new manager's (in terms of fixed effects) is roughly 0 (0.014) for all arrivals (including those not considered as events), reinforcing that manager arrivals are quasi-random and not made on the basis of manager or store quality. The mean fixed effect for an 'above average' manager (i.e. someone who raises the managerial quality at the store) is 0.511. store and month Fixed Effects are included. The control group includes never-treated stores.

Next, we quantify the relative importance of managers and the overall quality of workers. We find that a 1 SD increase in the quality of a manager has roughly 1.4x the

<sup>54</sup>The arrival of a new manager is coded as an 'event' if the store manager has a higher estimated fixed effect than the mean fixed effect of the managers who preceded the new manager's arrival.

<sup>55</sup>The mean fixed effect for an 'above average' manager (i.e. someone who raises the managerial quality at the store) is 0.511. To find the impact of a 'good manager' (i.e. someone who is 1 SD above the average) we scale the event study estimates by  $1/0.511$ . Average monthly sales in the stores we analyze is \$1.34 million USD.

impact on sales as a 1 SD increase in the average quality of workers at the store <sup>56</sup> This is comparable to our findings in the lab, where we estimate the effect of a 1 SD improvement in manager performance is 1.1x larger than the effect of a 1 SD improvement in the productive skills of workers.

We then turn to the important question of whether the predictors of managerial performance in the lab also predict performance in the field. This is an important test of external validity, in that we hypothesize similar underlying skills should drive success in both contexts (Findley et al., 2021). We find that this is indeed the case. We regress manager fixed effects on the three broad measures of managerial skill we used in the experimental study. The results are presented in Table 6. The best predictor of performance is the measure of economic decision-making: a one SD increase in economic decision-making skill is associated with an increase in manager fixed effects of 0.19 SD (this translates to a mean increase of around 4.9% of monthly sales, equating to 794k USD per annum, per store). We find that the association between economic decision-making and managerial performance is robust to controls for age, gender and work experience, mirroring the results from the lab.

Table 6 also provides information about the extent to which observable characteristics of managers (skill measures and demographic factors) explain managerial performance. Column 6 shows that around 24% of the variation in manager fixed effects is explained by skill measures, age, gender and experience ( $R^2 = 0.238$ ). This result is very close to findings from the lab experiment: the same vector of ex-ante observable characteristics explain 20% of variation in causally-identified managerial contributions (see column 11 of table B.2). The similarity suggests that despite the many differences between the lab and field settings we study, there is significant overlap in the role played by skills and demographic factors in managerial performance in both contexts.

### 4.3 Quantifying the impact of managerial selection in the field

Last, we quantify the economic impact of skill-based hiring in the field. This is the field analogue of the experimental analysis comparing the effect of different selection mechanisms on typical managerial performance. In this case we extend the analysis, to consider the impact not just on management quality, but on sales. In particular, we consider the impact on sales for a store that shifts from its current hiring policy to

---

<sup>56</sup>Unfortunately, unlike in the experiment reported in Section 3, we do not have a direct measure of workers' production skills. To contextualize the impact of managerial skills relative to worker skills, we estimate naive fixed effects for all workers in our sample, in terms of log sales. We then take the mean worker fixed effect in each store at a given time to obtain a measure of the average worker quality. The contribution of a 1 SD greater average worker quality to sales is estimated as 18%, as compared to the 25% obtained for managers from the event study (a ratio of 1.4x).

Table 6: Which skills predict manager performance in the field?

|                                 | (1)<br>Managers' FEs:<br>Demog. Controls | (2)<br>Managers' FEs:<br>Demog. Controls | (3)<br>Managers' FEs:<br>Demog. Controls | (4)<br>Managers' FEs:<br>Demog. Controls | (5)<br>Managers' FEs:<br>Demog. Controls |
|---------------------------------|--|--|--|--|--|
| AG score                        | 0.192**<br>(0.0691)                      |  |  | 0.242**<br>(0.0968)                      | 0.273**<br>(0.101)                       |
| Ravens' Test score              |  | -0.0327<br>(0.0903)                      |  | -0.0525<br>(0.119)                       | -0.0890<br>(0.137)                       |
| Reading the Mind<br>in the Eyes |  |  | -0.152<br>(0.0945)                       | -0.184*<br>(0.0918)                      | -0.180*<br>(0.0914)                      |
| Female                          |  |  |  |  | 0.161*<br>(0.0902)                       |
| Age                             |  |  |  |  | -0.151<br>(0.165)                        |
| Experience                      |  |  |  |  | 0.0466<br>(0.155)                        |
| Observations                    | 94                                       | 94                                       | 94                                       | 94                                       | 94                                       |
| R-squared                       | 0.177                                    | 0.145                                    | 0.164                                    | 0.210                                    | 0.238                                    |
| FEs                             | C set                                    | C set                                    | C set                                    | C set                                    | C set                                    |
| Cluster                         | C set                                    | C set                                    | C set                                    | C set                                    | C set                                    |

**Note:** Table 6 presents the correlation between the estimated manager's FE (from equation 4) and the Assignment Game score, Ravens' test, and Reading the Mind in the Eyes test. To recover the manager FEs, we estimate a two-way fixed effect model (AKM), regressing log monthly store sales on both store and manager fixed effects, controlling for time fixed effects and average demographic characteristics at the store level including Gender, marital status, type of contract, unionized, age, experience. This regression is a cross section at the manager level. All RHS variables are in standard deviations. Sample: managers who answered the Assignment Game, excluding connected set 0. Standard errors reported in parentheses are clustered at the connected set level. \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .

one in which managers are hired according to economic decision-making skills (the best predictor of managerial performance in the lab and the field).

To calculate the potential impact of this policy relative to business-as-usual we mirror the approach in section 3.4. First, we rank managers according to their scores on the Assignment Game (i.e. the test of economic decision-making). We then consider the average managerial performance of the top tercile in this group.<sup>57</sup> Last we translate this effect into the impact on sales by using the event study analysis that links a one SD improvement in management quality to store-level sales. We estimate that if a store were to fill a managerial vacancy by filtering candidates according to their economic decision-making scores, this would improve sales by 4.6% percent. This estimate is likely conservative as it only considers the skill profile of existing store managers. Evidence from our lab experiment, and from Benson et al. (2019), suggests that these managers may be far from optimally selected. We therefore conduct a complementary analysis that is calibrated using data from the lab experiment. We start with cleanly identified estimates from our experiment about the impact of filling a vacancy with a skills-based hire. We combine these estimates with data from the field that quantify the value of good managers in terms of store-level sales. Overall, we estimate that shifting to an approach which only considered candidates from the top third of scorers on the

<sup>57</sup>This threshold was chosen to mirror the experimental analysis. We also consider a regime where we focus on the top 9% of managers in terms of their assignment game scores, as 9% of staff in the field settings are managers. This tighter threshold implies that shifting from business-as-usual to a new policy of hiring according to Assignment Game scores would increase sales by %7.9.

assignment game would increase average monthly sales by 16%.<sup>58</sup>

## 5 Mechanisms

This section examines three potential mechanisms by which good managers improve team performance: monitoring; allocating tasks according to comparative advantage; and motivating workers to be more productive. In the lab experiment we are able to compare the relative importance of these three channels. We find that monitoring and motivation are stronger predictors of manager contributions, though allocative skill plays a substantive role too.

We present complementary evidence from the field, showing that good managers reduce monitoring errors that lead to product stockouts.<sup>59</sup>

### 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 tasks that don't contribute to group performance. 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.<sup>60</sup>

We arbitrarily define a module score as being "substantially greater" than the team score if it is 10 points higher than the minimum 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 than the minimum module score.

Over the course of the experiment, 13% of groups finished the task with at least one person working on a module that was not contributing to group success. Failures in

---

<sup>58</sup>The magnitude of this effect is roughly half the size of the counterfactual estimated in Benson et al. (2019), who report that candidates selected using an optimal combination of observable characteristics increase sales on average by 30%.

<sup>59</sup>The field analysis was not pre-registered, and is largely exploratory in nature.

<sup>60</sup>It may be the case that managers are worried that a worker who is unskilled may reduce the score on the bottlenecked module. In these instances it may make sense for them to allocate the worker to a module where they do no harm. This happens in roughly 2% of groups and we do not code such instances as monitoring errors. We thank an anonymous reviewer for pointing out this possibility.

monitoring were strongly related to overall manager contributions ( $\hat{\alpha}_i$ ). The bivariate correlation between monitoring errors and manager performance is -0.34 ( $p < 0.001$ ,  $n = 186$ ). A manager 1 SD above average reduced the error rate from 13.3% to 7.1%. In other words, good managers had roughly half the rate of monitoring errors.

## 5.2 Allocating 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 sample to groups who *initially* assigned one person to each of the three modules.<sup>61</sup> 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 unambiguously 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 that a manager finds the optimal starting assignment is positively associated with overall manager performance ( $\rho = 0.18$ ,  $p = 0.02$ ). 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 0.50 SD 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.

## 5.3 Motivating Workers to Exert Effort

The final mechanism we examine is worker motivation. We indirectly measure how motivated each worker is by examining variation in worker productivity. Let  $Y_{ik[m]}$  be the rate at which worker  $i$  correctly solves problems on domain  $k$  when they're working for manager  $m$ .

We fit a simple model to help understand variation in each worker's productivity across managers and modules:

$$Y_{ik[m]} = \nu_m + \delta_k + \beta X_{ik} + e_{ik} \quad (5)$$

Where  $\nu_m$  represents manager fixed effects,  $\delta_k$  are domain fixed effects for the three

---

<sup>61</sup>90% of groups used this strategy.

module-types (analogical; numerical; spatial);  $X_{ik}$  represents worker  $i$ 's estimated solo ability to solve problems in domain  $k$  measured in the pre-test.<sup>62</sup> There are two things to note about equation (5). First, because managers are randomly assigned to teams, the manager fixed effects  $\nu_m$  represent the systematic causal impact manager  $m$  has on their workers productivity (across rounds and modules). As we control for how good workers are at solving different problems, and as they are working alone on these problems, we characterise systematic changes in productivity as the manager's effect on worker motivation. Second, note that the manager fixed effects in equation (5) differ in an important way from the 'manager effects' we estimate in the main analysis. The overall manager effects ( $\alpha_m$ ) focus on how the presence of manager  $m$  affects *overall group performance* – not merely the rate at which worker  $i$  completes tasks ( $Y_{ik[m]}$ ). In principle, a manager could be an excellent motivator ( $\nu_m \gg 0$ ) but have a negative overall impact on their team ( $\alpha_m < 0$ ) if the manager fails to use comparative advantage to allocate worker tasks and/or they make monitoring errors whereby workers spend time solving problems that do not contribute to group performance. Overall, we find that 'motivational' fixed effects ( $\nu_m$ ) are an important driver of managerial performance ( $\hat{\rho} = 0.39$ ,  $p < 0.001$ ,  $n = 185$ ). Our finding that good managers were likely to be good motivators is plausible given that the tasks workers were assigned were cognitively demanding, and there were no financial incentives for workers to exert effort (as noted in Section 2).

## 5.4 Which mechanisms matter most in the experiment?

To understand the relative importance of the three channels, Table 7 regresses estimated manager effects  $\hat{\alpha}_i$  against three manager-level parameters: i) the rate at which manager  $i$  makes monitoring errors; ii) the rate at which managers start their group with an optimal allocation; iii) the 'motivation' fixed effect described in equation (5), measuring the causal impact that managers have on the productivity of their workers. All variables are standardized such that the table reports the impact of a one SD change. We note that the first two of these channels are both fundamentally about allocation (i.e. deciding who does what); the third channel focuses on the question of how productive a worker is once they've been assigned a specific task.

We find that all three mechanisms independently contribute to managerial contributions (column 5) and that the motivational and monitoring channels are the strongest pre-

---

<sup>62</sup>One possibility is that manager  $m$  is skilled at domain  $k$  and helps worker  $i$  solve problems. In this case, the mechanism under investigation would be better characterized as 'team support' than 'motivation'. To account for this, in a robustness check we included  $X_{km}$  in equation (5) to control for the managers ability to solve problems in domain  $k$ . This variable has no impact on  $Y_{ik[m]}$  suggesting that 'managers helping workers solve problems' is not an important channel in our analysis. We therefore excluded  $X_{mk}$  from equation (5) for simplicity.

Table 7: Which mechanisms predict managerial performance in the lab?

|                    | <i>Dependent variable: <math>\alpha_i</math></i> |                      |                     |                      |                      |
|--------------------|--|----------------------|---------------------|----------------------|----------------------|
|                    | (1)  | (2)                  | (3)                 | (4)                  | (5)                  |
| Initial allocation | 0.204***<br>(0.072)                              |                      |                     | 0.178***<br>(0.068)  | 0.136**<br>(0.064)   |
| Monitoring errors  |  | -0.335***<br>(0.069) |                     | -0.321***<br>(0.068) | -0.348***<br>(0.064) |
| Worker Motivation  |  |                      | 0.391***<br>(0.068) |                      | 0.383***<br>(0.064)  |
| Observations       | 185  | 185                  | 185                 | 185                  | 185                  |
| $R^2$              | 0.042  | 0.113                | 0.154               | 0.145                | 0.292                |
| Adjusted $R^2$     | 0.037  | 0.109                | 0.150               | 0.136                | 0.281                |

*Notes:* All predictors are measured at the level of individual managers, and are standardized across the sample to have SD=1, mean=0. 'Initial allocation' measures the frequency with which a manager finds the optimal initial allocation; 'Monitoring errors' is based on the final allocation of workers to tasks and measures the number of times a worker was expending effort on a module that could not improve group performance; 'Worker motivation' is measured by each manager's causal effect on the rate at which individual workers solve puzzles on different modules (controlling for worker differences in module skill which were assessed before the group task). standard errors in parentheses. \*\*\* $p < 0.01$ , \*\* $p < 0.05$ , \* $p < 0.1$ .

dictors of overall manager performance. A one SD increase in manager's performance in terms of motivation improves overall managerial performance by 0.39 SD. These estimates are virtually unchanged in a model that simultaneously controls for all three variables (column 5).

Table 7 also allows for a comparison of the combined impact of 'decision-making' (allocating and monitoring, presented in column 4) vs 'productivity' (motivation, presented in column 3). Both of these broad channels have similar explanatory power, with each independently explaining roughly 15% of the variance in managerial contributions.

## 5.5 Mechanisms in the field

Finally, we test whether the mechanisms we identify in the lab are present in the field. We find suggestive evidence in support of the hypothesis that *monitoring* is an important channel for managerial performance. Specifically, we examine 'stockouts', a situation in which managers fail to monitor stock levels leading to products becoming unavailable.

We find that good managers reduce the rate of stockouts. We conduct another event study analysis, focusing now on the impact that a manager arrival event (defined once again as the arrival of a manager who raises the managerial quality of the store) have on the rate of stockouts. We find that the arrival of a higher quality manager reduces the proportion of products out of stock by roughly .3pp, or 4.4% of the mean.



Finally, we test whether managers with strong economic decision-making skills are more likely to successfully monitor stock, identify when items are running low and proactively avoid stockouts. We find a strong marginal correlation between assignment game score and the rate of stockouts in stores, a 1 SD increase in managers' assignment game score being associated with a 3.5 ppt reduction in stockouts,  $SE = 0.16$ ).

Figure 6: Impact of 'above-average' managers on *Stockouts*: Event Study

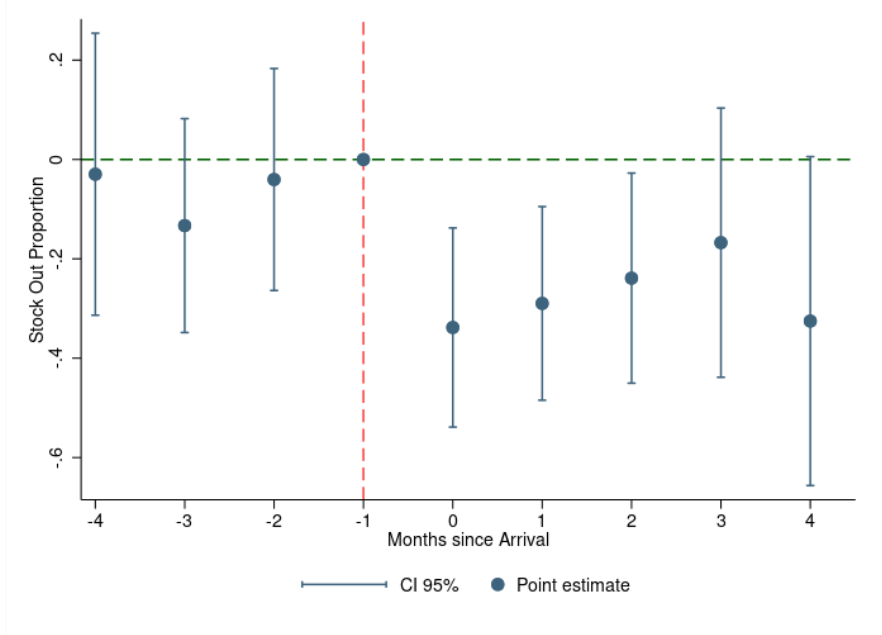


Figure 6 shows the results from estimating an event study design using the arrival of a manager with a log sales fixed effect bigger than the mean store log sale fixed effect across pre-existing managers in the store. The dependent variable is the proportion of items for which there is zero stock at the end of each month. Manager fixed effects are estimated using 4. Standard errors are clustered at the store level. The average proportion of products out of stock in a store month is 6.83%.

## 6 Discussion and Conclusion

This paper develops a novel, skills-based approach to prospectively identifying good managers. In a lab setting, we repeatedly randomly assign managers to teams of people who work on a collaborative production task. We estimate the casual contribution of the manager to group performance after conditioning on each team's endowment of production skills. Over multiple random assignments some managers consistently cause their teams to exceed predicted performance. We find that manager quality matters about as much as the *overall* productive capacity of workers.

We then explore the characteristics that predict our causal measure of managerial 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. Exploratory analysis suggests that this may be partly due to overconfidence. People who want to be in charge are significantly more likely to overrate their performance compared to their objective contributions. Similarly, self-promoted managers report higher social skills, but do worse on a widely used skill-based test of emotional perceptiveness. Taken together, these characteristics may reduce managerial effectiveness by undervaluing workers' perspectives and skills.

Our experimental measure of manager quality relies on a purpose-built collaborative task which requires managers to coordinate, monitor and motivate workers (Metcalf et al., 2023a; Minni, 2023; Fenizia, 2022; Adhvaryu et al., 2023; Benson and Shaw, 2025; Dessein et al., 2024). While these are crucial aspects of many managerial roles our approach does not capture other important managerial skills, especially those associated with long-term, repeated interactions, such as relationship management and conflict resolution.<sup>63</sup>

This raises the question of how well our skills-based approach predicts managerial success in field settings. We conduct two studies to provide initial evidence of external validity. First, we find that experimental participants who make strong causal contributions as managers in our lab setting are more likely to receive early-career promotions in the labor-market. Second, we study middle managers inside a large retail firm. We find that differences in managerial quality are economically meaningful (a one SD improvement is associated with a 25 percent increase in sales) and that the strongest predictor of managerial performance in the lab – a test of economic decision making – is also the strongest predictor in the field.

What practical steps could firms take in response to this evidence? In general, firms need to navigate a trade-off between the cost of acquiring information about candidates, and the predictive power of the information. At the low-cost end of this spectrum, firms may benefit from including a short assessment of economic decision-making skills in their hiring process (Caplin et al., 2024). Findings from both the lab and the field suggest that this test is associated with managerial performance. The full test is relatively short (around 15 minutes) and a short-form version of the test taking 5-8 minutes has also been validated in a large sample of Danish workers (Caplin et al., 2024). Moreover, the relatively poor performance of self-promoted managers in our study suggests that firms may not want to be limited to candidates who put themselves forward. Instead, they may benefit from proactively assessing potential candidates – for example, by fielding the Assignment Game among frontline employees. As short, publicly available

---

<sup>63</sup>Our approach also focuses on middle managers and does not capture skills that are more important at higher levels of leadership, such as creating and sustaining a positive company culture (Antonakis et al., 2022).

assessments are available, we believe that this approach to identifying candidates with strong managerial skills may be feasible even for cost-constrained organizations.

Firms with greater scale could consider adopting our approach of repeatedly randomizing (prospective) managers to multiple teams. Large organizations already use group-based assessments in their hiring processes. For example, some candidates applying to the UK Civil Service complete a 5-hour assessment battery that includes a group-based test of teamwork and communication (Assessment Centre HQ, 2025). We believe it would be possible to modify existing group assessments to mirror the repeated randomization design described in this paper. For organizations that have not previously completed group-based assessments, instead of developing their own task they could use the Collaborative Production Task (the materials for which are freely available).

We acknowledge that group-based assessments are challenging, and it may be the case that firms are limited by logistical constraints i.e. they may not be able to mimic our experimental design and simultaneously test 12-18 candidates at a time. In such cases, we note that it may be possible to use AI agents to simulate the role of human workers in a repeated-randomization design. With such a setup, candidates would complete a 1 hour, single-person assessment in which a leader is repeatedly assigned to teams of AI agents to solve collaborative puzzles. Recent evidence suggests that estimating leadership skills using this approach generates results that are remarkably similar to results obtained from repeatedly randomizing leaders to multiple groups of humans (Weidmann et al., 2025).<sup>64</sup>

In conclusion, our findings emphasize the importance of good management and suggest that skill measures are better predictors of manager performance than personality traits or preferences to be in a leadership position. This is important because preferences for leadership and qualities like extraversion and self-confidence increase the probability of promotion in many workplaces. Our results suggest that firms who proactively engage the widest possible set of candidates and screen for skills such as economic decision making, could see substantial improvements in managerial quality and team performance.

---

<sup>64</sup>A final option that organizations could consider is to have an extensive probation period, during which time manager candidates are randomly assigned to multiple teams within a workplace. In contexts where team performance can be measured this would result in clean causal estimates of manager effects. This option would generate the most ecologically valid and predictive estimates of managerial contribution - but would be prohibitively costly and disruptive.

## References

- ABOWD, J. M., R. H. CREECY, AND F. KRAMARZ (2002): “Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data,” Longitudinal Employer-Household Dynamics Technical Papers.
- ABOWD, J. M., F. KRAMARZ, AND D. N. MARGOLIS (1999): “High wage workers and high wage firms,” Econometrica, 67, 251–333.
- ADHVARYU, A., V. BASSI, A. NYSHADHAM, AND J. TAMAYO (2024): “No line left behind: Assortative matching inside the firm,” Review of Economics and Statistics, 1–45.
- ADHVARYU, A., A. NYSHADHAM, AND J. TAMAYO (2023): “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.
- ALTONJI, J. G. AND C. R. PIERRET (2001): “Employer Learning and Statistical Discrimination,” The Quarterly Journal of Economics, 116, 313–350.
- ANDREWS, M., L. GILL, T. SCHANK, AND R. UPWARD (2008): “High Wage Workers and Low Wage Firms: Negative Assortative Matching or Limited Mobility Bias?” Journal of the Royal Statistical Society, 171, 673–697.
- ANTONAKIS, J., G. D’ADDA, R. WEBER, AND C. ZEHNDER (2022): ““Just words? Just speeches?” On the economic value of charismatic leadership,” Management Science, 68, 6355–6381.
- ASSESSMENT CENTRE HQ (2025): “Civil Service Fast Stream Assessment Centre Guide,” <https://www.assessmentcentrehq.com/civil-service-fast-stream-assessment-centre/>.
- 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.
- BASS, B. M. AND B. J. AVOLIO (2000): MLQ Multifactor Leadership Questionnaire, Redwood City: Mind Garden.

- BEENEN, G., S. PICHLER, B. LIVINGSTON, AND R. RIGGIO (2021): “The good manager: Development and validation of the managerial interpersonal skills scale,” Frontiers in Psychology, 12, 631390.
- BELL, S. (2007): “Deep-level composition variables as predictors of team performance: A meta-analysis,” Journal of Applied Psychology, 92, 595.
- BENDER, S., N. BLOOM, D. CARD, J. VAN REENEN, AND S. WOLTER (2018): “Management practices, workforce selection, and productivity,” Journal of Labor Economics, 36, S371–S409.
- BENSON, A., D. LI, AND K. SHUE (2019): “Promotions and the Peter Principle\*,” The Quarterly Journal of Economics, 134, 2085–2134.
- BENSON, A. M. AND K. L. SHAW (2025): “What Do Managers Do? An Economist’s Perspective,” NBER, working Paper No. w33431.
- BERGER, J., M. OSTERLOH, K. ROST, AND T. EHLMANN (2020): “How to Prevent Leadership Hubris? Comparing Competitive Selections, Lotteries, and Their Combination,” The Leadership Quarterly, 31, 101388.
- BERTRAND, M. AND A. SCHOAR (2003): “Managing with style: The effect of managers on firm policies,” The Quarterly Journal of Economics, 118, 1169–1208.
- 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.
- (2010): “New approaches to surveying organizations,” American Economic Review, 100, 105–109.
- 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.

- BRASS, D. J. (1984): “Being in the Right Place: A Structural Analysis of Individual Influence in an Organization,” Administrative Science Quarterly, 518–539.
- BRUHN, M., D. KARLAN, AND A. SCHOAR (2018): “The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico,” Journal of Political Economy, 126.
- 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.
- CARD, D., J. HEINING, AND P. KLINE (2013): “Workplace heterogeneity and the rise of West German wage inequality,” The Quarterly journal of economics, 128, 967–1015.
- 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.
- CHAN, D. AND N. SCHMITT (2002): “Situational Judgment and Job Performance,” Human Performance - HUM PERFORM, 15, 233–254.
- DESSEIN, W., D. H.-F. LO, R. SHANGGUAN, AND H. OWAN (2024): “The Management of Knowledge Work,” Tech. Rep. 24-E-044, RIETI.
- DOHMEN, T., A. FALK, D. HUFFMAN, AND U. SUNDE (2018): “On the relationship between cognitive ability and risk preference,” Journal of Economic Perspectives, 32, 115–134.
- 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.
- DORN, D., F. SCHONER, M. SEEBACHER, L. SIMON, AND L. WOESSMANN (2024): “Multidimensional Skills as a Measure of Human Capital: Evidence from LinkedIn Profiles,” arXiv preprint arXiv:2409.18638.
- 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.

- ERKAL, N., L. GANGADHARAN, AND E. XIAO (2022): “Leadership selection: Can changing the default break the glass ceiling?” The Leadership Quarterly, 33, 101563.
- 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.
- EXLEY, C. L. AND J. B. KESSLER (2022): “The gender gap in self-promotion,” The Quarterly Journal of Economics, 137, 1345–1381.
- 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. R., D. C. TREADWAY, R. W. KOLODINSKY, W. A. HOCHWARTER, C. J. KACMAR, C. DOUGLAS, AND D. D. FRINK (2005): “Development and validation of the political skill inventory,” Journal of Management, 31, 126–152.
- FINDLEY, M. G., K. KIKUTA, AND M. DENLY (2021): “External validity,” Annual review of political science, 24, 365–393.
- GIARDILI, S., K. RAMDAS, AND J. W. WILLIAMS (2022): “Leadership and productivity: a study of US automobile assembly plants,” Management Science.
- GIORCELLI, M. (2019): “The Long-Term Effects of Management and Technology Transfers,” American Economic Review, 109, 121–152.
- GOODALL, A. H. AND M. OSTERLOH (2015): “Women have to enter the leadership race to win: Using random selection to increase the supply of women into senior positions,” Tech. rep., IZA Discussion Papers.
- 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.
- GOSNELL, G. K., J. LIST, AND R. METCALFE (2020): “The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains,” Journal of Political Economy, 128, 1195–1233.

- GÜTH, W., M. V. LEVATI, M. SUTTER, AND E. VAN DER HEIJDEN (2007): “Leadership and Cooperation in Public Goods Experiments,” Journal of Public Economics, 91, 1023–1042.
- HAEGELE, I. (2024): “The broken rung: Gender and the leadership gap,” arXiv preprint arXiv:2404.07750.
- HAIER, R. J., R. COLOM, D. H. SCHROEDER, C. A. CONDON, C. TANG, E. EAVES, AND K. HEAD (2009): “Gray matter and intelligence factors: Is there a neuro-g?” Intelligence, 37, 136–144.
- HECK, P. R., A. K. BROWN, AND C. F. CHABRIS (2024): “The Social Sensing Hypothesis,” Manuscript.
- HIRSHLEIFER, J. (1983): “From Weakest-Link to Best-Shot: The Voluntary Provision of Public Goods,” Public Choice, 41, 371–386.
- HOFFMAN, M., L. B. KAHN, AND D. LI (2017): “Discretion in Hiring\*,” The Quarterly Journal of Economics, 133, 765–800.
- HOFFMAN, M. AND C. T. STANTON (2024): “People, Practices, and Productivity: A Review of New Advances in Personnel Economics,” NBER Working Paper, 32849.
- HUFFMAN, D., C. RAYMOND, AND J. SHVETS (2022): “Persistent overconfidence and biased memory: Evidence from managers,” American Economic Review, 112, 3141–3175.
- HURTZ, G. AND J. DONOVAN (2000): “Personality and Job Performance: The Big Five Revisited,” The Journal of applied psychology, 85, 869–79.
- JANKE, K., C. PROPPER, AND R. SADUN (2019): “The impact of CEOs in the public sector: Evidence from the English NHS,” Tech. Rep. w25853, National Bureau of Economic Research.
- JAVALAGI, A. A., D. A. NEWMAN, AND M. LI (2024): “Personality and leadership: Meta-analytic review of cross-cultural moderation, behavioral mediation, and honesty-humility,” Journal of Applied Psychology.
- JENSEN, U. T., L. B. ANDERSEN, L. L. BRO, A. BØLLINGTOFT, T. L. M. ERIKSEN, A.-L. HOLTEN, C. B. JACOBSEN, J. LADENBURG, P. A. NIELSEN, H. H. SALOMONSEN, N. WESTERGÅRD-NIELSEN, AND A. WÜRTZ (2019): “Conceptualizing and Measuring Transformational and Transactional Leadership,” Administration & Society, 51, 3–33.



- JUDGE, T. A., J. E. BONO, R. ILIES, AND M. W. GERHARDT (2002): “Personality and leadership: A qualitative and quantitative review,” Journal of Applied Psychology, 87, 765–780.
- JUDGE, T. A., A. E. COLBERT, AND R. ILIES (2004): “Intelligence and leadership: A quantitative review and test of theoretical propositions,” Journal of Applied Psychology, 89, 542–552.
- KAHNEMAN, D. AND G. KLEIN (2009): “Conditions for intuitive expertise: a failure to disagree,” American Psychologist, 64, 515.
- KLINE, P., R. SAGGIO, AND M. SØLVSTEN (2020): “Leave-out Estimation of Variance Components,” Econometrica : journal of the Econometric Society, 88, 1859–1898.
- LARSON, J. (2013): “In Search of Synergy in Small Group Performance,” In Search of Synergy: In Small Group Performance, 1–427.
- LAZEAR, E. P., K. L. SHAW, AND C. T. 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 (2023a): “Managers and Productivity in Retail,” Working Paper 2023-64, University of Chicago, Becker Friedman Institute for Economics.
- METCALFE, R. D., A. B. SOLLACI, AND C. SYVERSON (2023b): “Managers and Productivity in Retail,” Working Paper 31192, National Bureau of Economic Research.
- 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.
- PARKER, R. AND C. WEIR (2022): “Multiple secondary outcome analyses: precise interpretation is important,” Trials, 23.
- 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.

- 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.
- ROTEMBERG, J. J. AND G. SALONER (2000): “Visionaries, Managers, and Strategic Direction,” The RAND Journal of Economics, 31, 693–716.
- 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.
- TRADENTA, J. M., A. NEELIM, AND J. VECCI (2025): “Gender differences in the self-promotion of prosocial behaviour: exploring the female modesty constraint,” Experimental Economics, 1–21.
- VAN DEN STEEN, E. (2005): “Organizational Beliefs and Managerial Vision,” The Journal of Law, Economics, and Organization, 21, 256–283.
- VAN DICK, R., V. CIAMPA, S. LIANG, K. MOLTZEN, L. MONZANI, N. K. STEFFENS, AND S. A. HASLAM (2018): “Identity leadership going global: Validation of the Identity Leadership Inventory across 20 countries,” Journal of Occupational and Organizational Psychology, 91, 697–728.
- 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.
- WALUMBWA, F. O., B. J. AVOLIO, W. L. GARDNER, T. S. WERNING, AND S. J. PETERSON (2008): “Authentic Leadership: Development and Validation of a Theory-Based Measure,” Journal of Management, 34, 89–126.
- WEIDMANN, B. AND D. J. DEMING (2021): “Team Players: How Social Skills Improve Team Performance,” Econometrica, 89, 2637–2657.
- WEIDMANN, B., Y. XU, AND D. DEMING (2025): “Measuring Human Leadership Skills with AI Agents,” Tech. Rep. w33662, National Bureau of Economic Research.
- WELLS, K. (2020): “Who manages the firm matters: The incremental effect of individual managers on accounting quality,” The Accounting Review, 95, 365–384.

WHETZEL, D. L. AND M. A. MCDANIEL (2009): “Situational judgment tests: An overview of current research,” Human Resource Management Review, 19, 188–202, employee Selection at the Beginning of the 21st Century.







## A Appendix

## A.1 Individual measures

### A.1.1 Measuring participants ability to be productive when working alone

Before the group testing began, we assessed individuals' ability to solve problems on their own. This involved three tests (selected to match the type of problems that the groups would be asked to tackle). 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 A.7. 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 A.7). 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 A.7).

Figure A.7: Example items measuring individual 'productive skills'

|  |   |  |
|--|---|--|
| <p><b>A</b></p> <p>What is the missing number in the sequence:</p> <p>70   71   76   ?   82   83</p> <p> <input type="radio"/> 73<br/> <input type="radio"/> 75<br/> <input type="radio"/> 80<br/> <input type="radio"/> 77<br/> <input type="radio"/> 81         </p> <p><b>Numerical</b></p> | <p><b>B</b></p>  <p>Which of the following CANNOT be built from the above shape?</p> <div style="display: flex; justify-content: space-around;"> <div style="text-align: center;">  <p>A</p> </div> <div style="text-align: center;">  <p>B</p> </div> </div> <div style="display: flex; justify-content: space-around;"> <div style="text-align: center;">  <p>C</p> </div> <div style="text-align: center;">  <p>D</p> </div> <div style="text-align: center;">  <p>E</p> </div> </div> <p><b>Spatial</b></p> | <p><b>C</b></p> <p>MURAL is to WALL, as<br/>INSCRIPTION is to _____</p> <p>           a. plaque<br/>           b. dedication<br/>           c. Brush<br/>           d. floor         </p> <p><b>Analytical</b></p> |
|--|---|--|

**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).

### A.1.2 Broad measures of individual skill

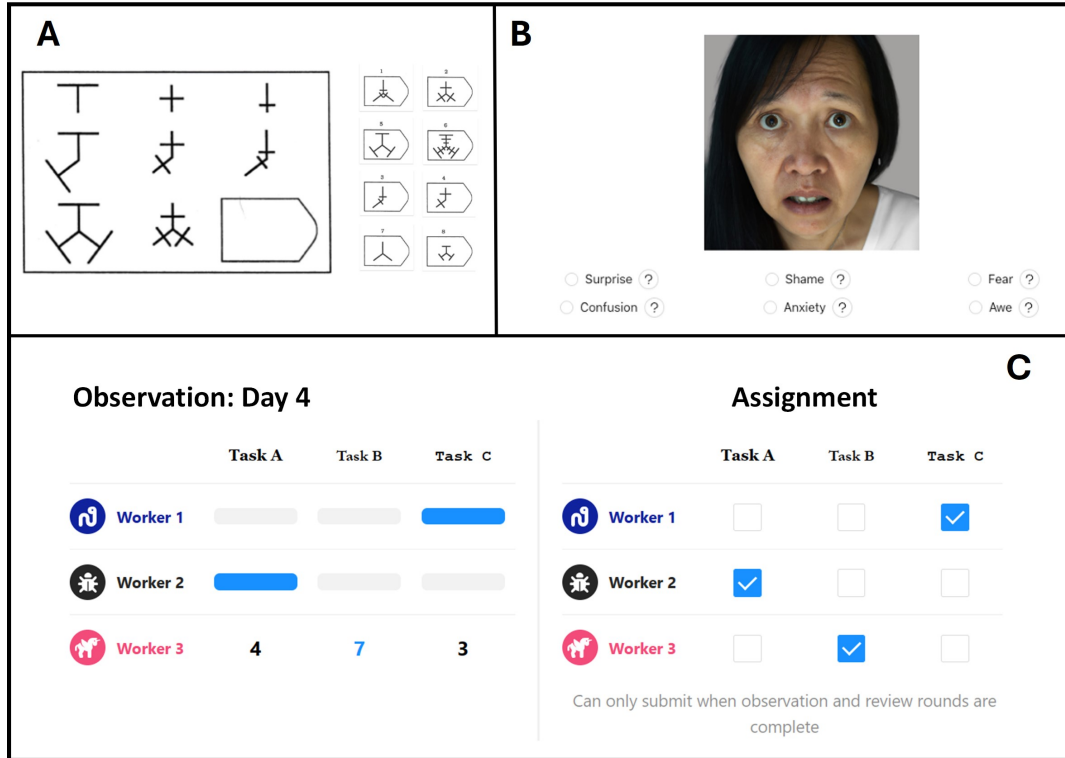
In addition to skills that were specifically chosen to match the demands of the group task, we measured three broad domains of individual skill: non-verbal fluid intelligence; emotional perceptiveness and economic decision-making skill. This section provides example problems from each of these assessments.

Fluid intelligence was measured using a set of Ravens Advanced Progressive Matrices.

## Supplementary material: For Online Appendix

An example problem is presented in Panel A, below. Emotional perceptiveness was measured using RMET (reading the mind in the eyes test, see Baron-Cohen et al., 2001). An example problem is presented in Panel B of Figure A.8.<sup>65</sup> Finally, economic decision-making, defined by Caplin et al., (2024) as the ability to make good resource allocation decisions, was assessed using the Assignment Game (see panel C of Figure A.8 for an example, and Caplin et al., (2024) for a detailed description).

Figure A.8: Example items from broad tests of individual skill



**Notes:** Panel A is an example item from Ravens, 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).

## A.2 Design Considerations for the Collaborative Production Task

The Collaborative Production Task was designed to satisfy four criteria. First, we sought a task that allowed for objective scoring to reduce measurement error. Second, we wanted collaboration to be essential for group success, a feature which is often lacking in group tasks (Larson, 2013). Third, we looked for a group task that had an individual analogue, which would allow us to control for differences in each team's

<sup>65</sup>Definitions of all the words were available via links to an online dictionary. We used a shortened version of the test, with 26 items rather than 36, as per Weidmann and Deming (2021).

endowment of individual productive skill. Finally, given our focus on managers, the task needed to have a clear and distinct role for managers which replicated real-world managerial skills (Brandts and Cooper, 2007; Güth et al., 2007; Sahin et al., 2015; Benson and Shaw, 2025). For this last criterion, we focused in particular on the recent work of Benson and Shaw (2025), which provides descriptive evidence on what managers actually do (from an economists perspective) using data from LinkedIn. (Benson and Shaw, 2025) emphasize two main managerial roles: ‘managers of people’ who ‘typically hire, retain, and motivate a larger set of individual contributors’ (p31) and ‘managers of projects’ who ‘typically monitor, coordinate, and allocate interdependent collaborators who work toward a common goal’ (p31). Ultimately, the Collaborative Production Task was designed to focus on four of these widespread managerial skills: motivating; monitoring; co-ordinating and allocating.

**Scoring Rule** In the Collaborative Production Task, groups are required to work on three question modules: numerical, spatial, and analytical reasoning. 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. We instituted a penalty for incorrect solutions to disincentivize guessing. Each person in the team, including the manager, works on their own computer trying to solve a different problem. Importantly, the manager decides who will be working on each module.

The overall team score is the *minimum* module score. As noted in the text, this is similar to the weakest link coordination game in which collaboration is essential for success. 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. An alternative we considered was 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. 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.

### A.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:

$$\hat{\Omega}_i = \frac{1}{\sum_g W_{ig}} \sum_g W_{ig} \hat{\epsilon}_g \quad (6)$$

## Supplementary material: For Online Appendix

A similar substitution can be made in equation (3) to estimate the typical worker effect, defined as  $\sigma_\Omega$ :

$$\begin{aligned}\hat{e}_{gi} &= \Omega_i + e_{gi} \\ \Omega_i &\sim N(0, \sigma_\Omega^2) \\ e_{gi} &\sim N(0, \sigma^2)\end{aligned}\tag{7}$$

In equation 7,  $\hat{e}_{gi}$  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.  $\Omega_i$  is a random worker effect for individual  $i$  on group  $g$ .

### A.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 the null hypothesis that managers have no impact on their teams after controlling for each team’s endowment of productive skill. Our 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.<sup>66</sup>

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  $\hat{\sigma}_\alpha^2$  to the simulated distribution under the null and calculate the frequency with which draws from the null distribution are greater than  $\hat{\sigma}_\alpha^2$ , i.e., we estimate  $Pr(\hat{\sigma}_{\alpha(NULL)}^2 > \hat{\sigma}_\alpha^2)$ . This is our p-value (Ernst, 2004).

## B Robustness Checks for Lab Experiment

### B.1 Leave-one-out validation

As an addition check for the robustness of the estimated manager and worker effects, we test whether estimates of  $\alpha_i$  and  $\Omega_i$  predict performance out-of-sample using a leave-

---

<sup>66</sup>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 (Venzone and Moolgavkar, 1988).



## Supplementary material: For Online Appendix

one-out (LOO) procedure. The procedure involves removing one of the four rounds of data then, using the remaining data, calculating the average causal contributions of individual managers ( $\alpha_i^{LOO}$ ) and workers ( $\Omega_i^{LOO}$ ). We then assess whether these leave-one-out estimates predict whether a group will be successful in the round of data we hold out.<sup>67</sup> We repeat this procedure for each of the four rounds and estimate the following model:

$$G_g = \beta_0 + \sum_i \hat{\alpha}_i^{LOO} M_{ig} + \sum_i \hat{\Omega}_i^{LOO} W_{ig} + \epsilon_g \quad (8)$$

The results are presented in Table B.1. 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.

Table B.1: Predicting manager and worker contributions out-of-sample

|   | Group performance in hold out round |                  |                  |                    |                     |
|---|-------------------------------------|------------------|------------------|--------------------|---------------------|
|   | Round 1<br>(1)                      | Round 2<br>(2)   | Round 3<br>(3)   | Round 4<br>(4)     | Overall<br>(5)      |
| Manager contribution LOO ( $\hat{\alpha}_i^{LOO}$ ) | 0.365***<br>(0.135)                 | 0.204<br>(0.163) | 0.151<br>(0.200) | 0.373**<br>(0.169) | 0.299***<br>(0.098) |
| Worker contribution LOO ( $\hat{\Omega}_i^{LOO}$ )  | 0.016<br>(0.135)                    | —                | 0.257<br>(0.200) | 0.113<br>(0.169)   | 0.129<br>(0.098)    |
| Observations  | 186                                 | 182              | 180              | 180                | 546                 |
| $R^2$   | 0.039                               | 0.009            | 0.013            | 0.032              | 0.021               |
| Adjusted $R^2$                                      | 0.028                               | 0.003            | 0.002            | 0.021              | 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  $\sigma_\Omega$  in this case is 0, meaning we cannot estimate worker effects. Standard errors are in parentheses and are calculated at the group level. Significance levels are denoted by: \*\*\*  $p < 0.01$ ; \*\*  $p < 0.05$ ; \*  $p < 0.1$

## B.2 Predictors of being a good manager in the lab

We find two reliable predictors of manager effects in our lab experiment: economic decision-making skill (as measured by the Assignment Game, discussed in Caplin et al.

<sup>67</sup>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.

## Supplementary material: For Online Appendix

(2024), and fluid IQ (as measured by Ravens Advanced Progressive Matrices). Table B.2 shows that these two predictors of management performance are robust to a wide range of controls, including demographics, education and work experience, and measures of emotional perceptiveness and personality.<sup>68</sup>

---

<sup>68</sup>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., 2023).

Table B.2: 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<br>(Assignment Game)                                | 0.295**<br>(0.131) | 0.342**<br>(0.139) | 0.388***<br>(0.150) | 0.365**<br>(0.150) | 0.400**<br>(0.158) |                    |                    |                     |                     |                     | 0.248<br>(0.184) |                     |
| Fluid intelligence<br>(Ravens)   |                    |                    |                     |                    |                    | 0.210**<br>(0.098) | 0.219**<br>(0.100) | 0.293***<br>(0.105) | 0.312***<br>(0.106) | 0.304***<br>(0.114) | 0.211<br>(0.133) |                     |
| Combined Decision-making<br>(AG+Ravens)                                      |                    |                    |                     |                    |                    |                    |                    |                     |                     |                     |                  | 0.358***<br>(0.118) |
| 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                   |
| Obs  | 90                 | 90                 | 86                  | 86                 | 86                 | 90                 | 90                 | 86                  | 86                  | 86                  | 86               | 86                  |
| R <sup>2</sup>   | 0.044              | 0.044              | 0.025               | 0.022              | 0.023              | 0.038              | 0.031              | 0.037               | 0.056               | 0.033               | 0.044            | 0.058               |

*Notes:* The dependent variable in these regressions is the manager contributions from our pre-specified model. We focus on data from the random arm of the experiment (n=90). Four participants have post-surveys missing, which reduces the sample to 86 for some specifications. Demographics include 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 Ravens test. Significance levels are denoted by: \*\*\*  $p < 0.01$ ; \*\*  $p < 0.05$ ; \*  $p < 0.1$ .

## C Labor-market experiences of experimental participants

Table C.1: Characteristics of participants in the follow-up sample

|                          | Follow-up<br>sample | No job<br>history | p-value          |
|--------------------------|---------------------|-------------------|------------------|
| Alpha                    | 0.131               | -0.091            | 0.154            |
| economic decision-making | -0.084              | 0.058             | 0.350            |
| Ravens                   | 0.047               | -0.032            | 0.597            |
| RMET                     | 0.043               | -0.029            | 0.615            |
| preference for manager   | 7.987               | 7.491             | 0.165            |
| self-promoted manager    | 0.47                | 0.50              | 0.659            |
| <b>Age (years)</b>       | <b>27.5</b>         | <b>23.5</b>       | <b>&lt;0.001</b> |
| Female (%)               | 46%                 | 44%               | 0.824            |
| N                        | 73                  | 111               |                  |

Notes: The table presents a comparison of *means* across people with full-time jobs recorded on LinkedIn (n=73 managers) and people with no job histories available (n=111). Among the people with no job histories, n=36 are students who are yet to enter the labor market; n=27 have a LinkedIn page, but no full-time job listing (these pages are often blank); and n=46 had no LinkedIn page. 'Alpha' is pre-registered experimental measure of management skill, defined in equation 2. Skill variables and personality variables are standardized to have mean=0, sd=1 (alpha; economic decision making; ravens, RMET, Big5). Preference for manager is on a scale from 1-10, where 10 is the strongest preference for management. Age is measured in years.

# Supplementary material: For Online Appendix

Table C.2: Regression Results for Promotions per Year

|                          | (1)                 | (2)                 | (3)                 | (4)                 | (5)                | (6)                 | (7)              | (8)                |
|--------------------------|---------------------|---------------------|---------------------|---------------------|--------------------|---------------------|------------------|--------------------|
| Alphas                   | 0.164***<br>(0.045) | 0.169***<br>(0.043) | 0.184***<br>(0.048) | 0.180***<br>(0.053) |                    |                     |                  | 0.152**<br>(0.060) |
| Task Skills              |                     | 0.097**<br>(0.043)  | 0.087*<br>(0.046)   | 0.105**<br>(0.050)  |                    |                     |                  | 0.094*<br>(0.054)  |
| Economic Decision-Making |                     |                     |                     |                     | 0.101**<br>(0.047) |                     |                  | 0.011<br>(0.068)   |
| Fluid IQ                 |                     |                     |                     |                     |                    | 0.125***<br>(0.047) |                  | 0.061<br>(0.059)   |
| Emotional Perceptiveness |                     |                     |                     |                     |                    |                     | 0.055<br>(0.049) | 0.041<br>(0.058)   |
| Demographics             |                     |                     | X                   | X                   |                    |                     |                  | X                  |
| Personality              |                     |                     |                     | X                   |                    |                     |                  | X                  |
| Observations             | 73                  | 73                  | 69                  | 69                  | 73                 | 72                  | 72               | 68                 |
| R <sup>2</sup>           | 0.161               | 0.216               | 0.279               | 0.320               | 0.061              | 0.091               | 0.018            | 0.353              |
| Adjusted R <sup>2</sup>  | 0.149               | 0.194               | 0.196               | 0.174               | 0.048              | 0.078               | 0.004            | 0.166              |

*Notes:* Analysis is at the participant level. The dependent variable is 'promotions per year'. Promotions are coded by two researchers, and are based on publicly-available LinkedIn data. See description in section 4.1 'Years in the labor market' is defined as the amount of time finishing undergraduate study and the date of data collection (January 2025), minus any periods when participants were studying full time. All skill measures are standardized to have mean=0 and sd=1. 'Task-skills' represent the average score on the underlying numerical/analytical/spatial tasks described in section 2. 'Personality' represents separate measures of the Big5: extraversion, conscientiousness, openness, agreeableness and emotional stability. 'Economic decision-making' is the score on the Assignment Game (Caplin et al. 2024). Fluid IQ is based on Ravens Advanced Progressive Matrices. Emotional perceptiveness is based on Reading The Mind in the Eyes (Baron-Cohen, 2001). See section 2 and appendix A for a full description of measures. Significance levels are denoted by: \*\*\*  $p < 0.01$ ; \*\*  $p < 0.05$ ; \*  $p < 0.1$ .

## D Other Tables and Figures for Retail

Table D.1: Descriptive statistics of middle managers

| Variable                                  | Mean   | Std. dev. |
|---|--------|-----------|
| Age                                       | 43.08  | 7.90      |
| Avg female proportion                     | 0.33   | 0.47      |
| Avg months of experience                  | 891.73 | 479.16    |
| Employment contract                       | 0.99   | 0.09      |
| Fixed-term contract of more than 3 months | 0.01   | 0.09      |
| Married                                   | 0.49   | 0.50      |
| Separate                                  | 0.05   | 0.22      |
| Single                                    | 0.28   | 0.45      |
| Free union                                | 0.19   | 0.39      |

**Note:** Table D.1 presents descriptive statistics for the sample of 225 middle managers of the retail company who participated in the assignment game. The table includes the average age, proportion of female respondents, average number of months of experience, and the shares of middle managers with an employment contract and with a fixed-term contract longer than three months. It also reports the distribution of respondents by marital status.

Table D.2: Manages vs workers quality at store-month level

| Variable  | (1)<br>Avg. log sales  |
|---|------------------------|
| Standardized average worker naive fixed effects | 0.187***<br>(0.0213)   |
| Constant  | 21.75***<br>(6.03e-06) |
| Observations                                    | 4,540                  |
| R squared                                       | 0.986                  |
| Connected set FE                                | YES                    |
| Cluster connected set FE                        | YES                    |
| Mean  | 21.75                  |
| SD  | 1.21                   |
| Avg. num. workers                               | 112.4                  |
| Avg. num. mngers                                | 4.5                    |

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

**Note:** Table D.2 presents the correlation between the logarithm of sales and worker fixed effects standardized at the connected set level. The panel is at the month level with 230 stores and 28 months. The RHS variables are the avg naive manager FE and average worker fixed effects at the manager-month level, both standardized. In estimating these naive fixed effects, we include the manager and worker fixed effects separately and control for store and time fixed effects. Trivial connected sets are not considered.

## Supplementary material: For Online Appendix

Table D.3: Manages vs workers quality at manager-store-month level

| Variables                    | (2)<br>log sale          | (4)<br>log sale          |
|------------------------------|--------------------------|--------------------------|
| Manager FEs:                 | 0.0850***                |                          |
| Naive FE regression          | (0.00805)                |                          |
| Standarized Workers FEs:     |                          | 0.180***                 |
| Naive FE regression          |                          | (0.0272)                 |
| Observations                 | 23,457                   | 2,376                    |
| R-squared                    | 0.987                    | 0.990                    |
| FEs                          | Store & Month<br>& C set | Store & Month<br>& C set |
| Cluster                      | Store & C Set            | Store & C Set            |
| Num. Clusters                | 210 & 54                 | 108 & 37                 |
| Dep. Mean (Thousand dollars) | 3,412                    | 1,725                    |
| Dep. Mean                    | 22.85                    | 22                       |
| Dep sd                       | 1.13                     | 1.2                      |

**Note:** Table D.3 presents the correlation between the logarithm of sales and manager and worker fixed effects standardized at the connected level. The panel is at the manager-store-month level. The RHS variables are the naive manager FE and average worker fixed effects at the manager-month level, both standardized. In estimating these naive fixed effects, we include the manager and worker fixed effects separately and control for store and time fixed effects. Trivial connected sets are not considered. The standard deviation for each avg FE is, Avg manager FE: 0.158; Avg worker FE: 0.147 Standard errors reported in parentheses are clustered at the store level.\*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

Table D.4: Variance Decomposition: Store-Manager AKM with Demographic Controls

|                             | Baseline<br>(1) | Andrews et al. (2008)<br>(2) | Leave-out Estimator<br>(3) |
|-----------------------------|-----------------|------------------------------|----------------------------|
| $\text{Var}(\theta)$        | 0.016           | 0.013                        | 0.010                      |
| $\text{Var}(\psi)$          | 0.056           | 0.0534                       | 0.050                      |
| $\text{Cov}(\psi, \theta)$  | -0.02           | 0.018                        | -0.015                     |
| $\text{Corr}(\psi, \theta)$ | -0.688          | 0.699                        | -0.685                     |

Table D.4 reports the estimates of equation (1) following the two-way fixed effects estimation procedure in Abowd et al. (1999).  $\theta$  corresponds to the Manager fixed effect;  $\psi$  to the Store fixed effect;  $\text{var}(\theta)$  and  $\text{var}(\psi)$ , are the variance of the manager and store fixed effects, respectively, and  $\text{Cov}(\psi, \theta)$  and  $\text{Corr}(\psi, \theta)$  the covariance and the correlation between the manager and store fixed effects. In column 2 we implement the Andrews et al. (2008) bias correction procedure to deal with limited mobility bias. In column 3, we allow for heteroskedasticity and implement the leave-out estimator proposed by Kline et al. (2020). These statistics are estimated only for the first connected set, which is the largest one. In line with limited mobility bias not being substantial in our setting, we show that our key findings are robust to all these types of corrections. We use Log sales after regressing it on time FEs and demographic controls. The variance of the residual log sale is 0.07.

## Supplementary material: For Online Appendix

Figure D.1: Card Event Study, Manager Store AKM

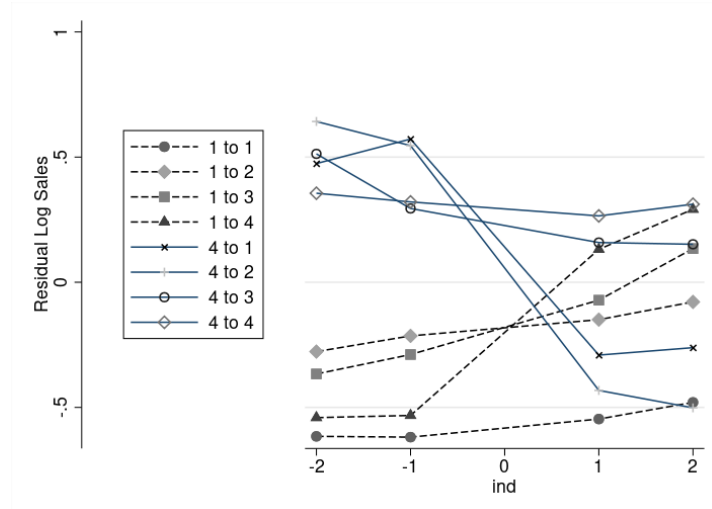


Figure D.1 assesses movers based on (i) quartiles of average Log Sales in their initial store and (ii) quartiles of the average Log Sales in the store where they moved to. The average Log Sale is computed over the entire sample period, and quartiles are calculated for each store. The graphical representation depicts the average residual in Log Sales of movers on the y-axis; the residual is computed for specific periods: from 3 to 4 (Period = -2) months and 1 to 2 months (Period = -1) before the move from the initial store, and 1 to 2 months (Period = 1) and 3 to 4 months (Period = 2) after the move to the new destination store, plotted on the x-axis. The analysis focuses on moves away from stores in the top quartile (lines in quartile 4) and stores in the bottom quartile (lines in quartile 1). To create the residual variable, we run a regression of the monthly Log sales of each store's on format (i.e. convenience, supermarkets, and hypermarkets FEs) and brand fixed effects. Then we predict the residuals and run the movers analysis.

Figure D.2: Card Event Study, Manager Store AKM

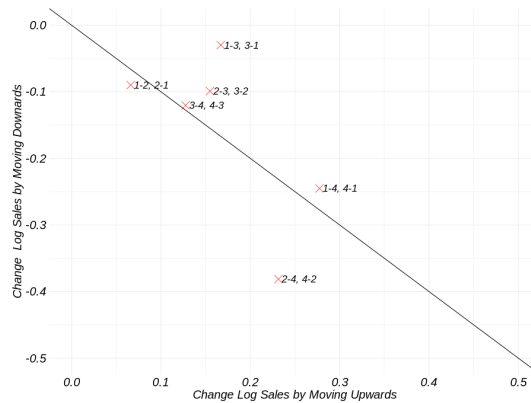


Figure D.2 ranks movers in terms of (i) quartiles of average Log Sales in their initial store and (ii) quartiles of the average Log Sales in the store where they moved to. The average Log Sales is computed over the entire sample period, and quartiles are calculated for each store. The Figure then plots the average change in residual Log Sales of movers from lines in quartile X to quartile Y, against the change in residual Log Sales for movers from lines in quartile Y to quartile X, plotted against the change for movers from lines in quartile 4 to quartile 2. The changes are calculated for the average residual Log Sale in the four months before the move and the four months after the move. The solid line corresponds to the 45-degree line. To create the residual variable, we run a regression of the monthly Log sales of each store's on format (i.e. convenience, supermarkets, and hypermarkets FEs) and brand fixed effects. Then, we predict the residuals and run the movers analysis.