Essays in the Organisational Economics of Management

Tom Schwantje

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Economics

October 2024

Abstract

This thesis comprises four chapters on the organisational economics of management.

The first chapter explores the use of rules by managers in making decisions. I develop a novel tool to measure the use of rules based on patterns in managerial decision-making, and implement this to survey senior Ethiopian human resource managers. I find that the use of rules is correlated in intuitive ways with a number of firm- and manager-level observable characteristics, validating the measurement approach. I find a robust, positive relationship between the use of formal (documented) rules and firm profitability, and that the use of such rules has substantial explanatory power for the differences in profitability across firms – beyond observable characteristics and measures of management practices. However, I find the adoption of informal (undocumented but commonly understood) rules appears to come at the cost of reduced resilience to economic shocks – specifically to COVID-19.

The second chapter – co-authored with Girum Abebe, Marcel Fafchamps, Michael Koelle and Simon Quinn – describes a novel matching experiment to examine the supply and demand for managerial traits. First, using studio responses to hypothetical management scenarios, analysed with a Bayesian hierarchical model, we describe heterogeneity in management styles among Ethiopian young professionals. We find four distinct managerial styles, and show labour market exposure and socioeconomic background correlate strongly with these managerial styles. Second, we show these recordings to firm managers, and show that firms have a clear preference for rule-oriented ("authoritative") managers who rely on formal policy and authority. Finally, leveraging a prior experiment, we show that these styles constitute a novel mechanism contributing to labour-market exclusion among individuals from disadvantaged socioeconomic backgrounds, and show a one-month management placement can effectively mitigate this labour-market exclusion.

The third chapter – co-authored with Girum Abebe, Siân Brooke, Tom Gole and Simon Quinn – covers a novel field experiment in Ethiopia, randomly varying the design of the assessments of a business plan competition across assessors. Using this experiment, we investigate how organisational values and social image concerns affect inclusivity in decision-making. We find that the treatments have no average effect on the inclusion of female entrepreneurs, but that the treatments appear to align individual identity with organisational values, improving decision quality and agreement among managers. We develop a stylised theoretical model to interpret these results.

The final chapter develops a mixed ordered logit mixture model to identify complementarities in management practices, leveraging data from the World Management Survey (WMS) to study what combinations of practices firms choose to employ. Using this method, I find significant variation across organisational structures and countries, and a clear structure in the combinations of practices that firms choose to employ.

Word Count: approximately 80,000

Acknowledgements

A DPhil can be—and was—a long road, and this thesis would not have been possible without the support of many people: my family, friends, and colleagues, all of whom I thank unreservedly.

First and foremost, I am deeply grateful to Simon Quinn, whose guidance shaped both the work presented here and my development as a researcher. He showed me that excellence and integrity in research can go hand in hand with generosity of time and support. I am also grateful to Romuald for his support in the later stages of the DPhil.

My sincere thanks go to the participants in the projects reported in this thesis for their time and contributions. I hope this work will, in some measure, benefit them and others. This research also owes a special debt to Gezahegn, whose tireless work in Addis made so much possible, and to Rose for steadfast support throughout—thank you.

I have been fortunate to learn from inspiring academics, including Christian, Daniela, Doug, Dennis, Girum, Ian, Johannes, Martin, Muhammad, Niklas, Sanjay, and Stephen. I am particularly grateful to Howard Smith for his insights and friendship during my time in Oxford, and to Chris Woodruff for detailed comments and advice on my work.

To my fellow students and friends—Juliette, Lisa, Rocco, Rosi, Sanghamitra, Vatsal, and Verena—thank you for your advice and friendship. To my Oxford friends beyond economics—a better group of friends than I could have wished for—thank you for the inspiration and the distractions. Oliver Chen, my big brother, for always setting the right example; Caro, for endless coffees and cinema trips; Flo, for being a constant inspiration and bringing chaos to my life; Liam, for a true example of dedication. Thank you to everyone at Teddy Hall, OULRC, OUTriC, and elsewhere—for everything.

To my family, bedankt voor alles. Lieve papa en mama, Olof en Linda, bedankt voor jullie oneindige steun, aanmoediging en vriendschap. Hopelijk heb ik uiteindelijk uit kunnen leggen wat ik nou eigenlijk doe. Lieve Hilde en Marit, mijn grote zussen: bedankt dat jullie er altijd voor me zijn.

Finally, I am grateful to the Oxford Department of Economics for financial support, which made this work possible.

Declaration of Authorship

I declare that this thesis includes collaborative work. The first and fourth chapters are solely my own work. The second chapter is co-authored with Girum Abebe, Marcel Fafchamps, Michael Koelle, and Simon Quinn, while the third chapter is co-authored with Girum Abebe, Siân Brooke, Tom Gole and Simon Quinn. For the second chapter, I led the project implementation and empirical analysis, and contributed to drafting the paper. For the third chapter, I led the project design, implementation, and analysis, and contributed to drafting the paper.

Tom Schwantje, October 2024

Contents

List of Tables

List of Figures

Introduction

This thesis consists of four chapters on the organisational economics of management, with the first three focusing on organisations in low-income countries. Each chapter focuses on a distinct topic, but the overarching theme is how firms tailor their organisational practices to their environment. The chapters in this thesis examine this theme by studying the use of rules by managers, the supply and demand for management traits, the importance of the institutional environment, and heterogeneity in management practices.

The first three chapters of this thesis draw on experimental and observational data from a unique sample of Ethiopian organisations, senior managers, and young professionals. This dataset contributes to the limited available data on large and medium-sized enterprises in low-income countries, which are fundamental for economic development, yet relatively underrepresented (Hsieh and Olken, 2014). An overarching theme in the findings based on this data is a desire for consistent, rules-based management – highlighting a stark difference in the choice of management practices between firms in high- and low-income countries.

1.1 Management as Rules: Evidence from an Adaptive Bayesian Questionnaire

The first chapter introduces a new measure of the use of rules by firm managers. Organisational decisions are often guided by rules that enforce consistency, standardisation, and fairness, potentially at the cost of reduced flexibility. This topic has received limited attention in the empirical literature in economics, at least in part because its high-dimensional nature does not lend itself to traditional economic tools. In this chapter, I develop a novel

tool to measure the use of organisational rules, and apply it to survey senior managers in Ethiopian firms.

Specifically, I develop a tool built on the principles of Bayesian optimisation (see, for example, Frazier, 2018), and adapt this to measure patterns in managerial decision-making. Based on these patterns, I then directly ask respondents whether they use specific rules that are consistent with these patterns, and whether these are formal, documented, or informal, undocumented but commonly understood, rules. I apply this tool to measure the use of rules among a sample of senior human resource managers in Ethiopian firms. I present the managers with a sequence of hypothetical scenarios, and measure the rules they use to make decisions in these scenarios.

I find that the elicited measure of rules correlates with a number of organisational characteristics, and with a set of firm management practices measured using more traditional 'Management and Organisational Practices Survey' (MOPS) questions. These correlations suggest a more general preference for more rule-based management practices among these firms (Bloom, Brynjolfsson, Foster, Jarmin, Patnaik, Saporta-Eksten, and Van Reenen, 2019). Next, I show that this measure helps explain heterogeneity in profitability across the sample of firms, even when controlling for a broad set of other firm covariates, and beyond the more traditional measure of management practices. In particular, the use of formal, documented rules is positively correlated with several measures of profitability. However, this higher performance appears to come at the cost of reduced resilience to economic shocks, indeed, firms with more rule-based management tend to have been significantly more negatively affected by COVID-19 in line with theoretical predictions (Li, Mukherjee, and Vasconcelos, 2022).

1.2 Lights, Camera, Transaction! Assessing Management Styles through Studio Vignettes

In the second chapter, the focus shifts to the supply and demand for managerial traits in Ethiopia. We conduct an incentivised matching experiment between firms and young professionals in Addis Ababa. This novel experimental design allows for direct elicitation

of both managerial skills and firms' preferences over those skills, enabling measurement and decomposition of heterogeneity in management styles and preferences. This chapter provides evidence on the heterogeneity and malleability of personality traits among young professionals and firms' demand for specific types of entry-level managers, and highlights a novel mechanism — and remedy — for intergenerational persistence of labour-market exclusion.

First, using a Bayesian hierarchical model, we find that there are four distinct latent types of managers among the young professionals. We interpret these, based on a seminal management paper (Goleman, 2000), as 'authoritative', 'affiliative', 'coercive', and 'timid'. The authoritative type is most distinct in terms of both the respondents' (superior) labour market outcomes and labour market experience. Strikingly, our findings suggest, both through correlational and causal evidence, that these managerial styles can be shaped by labour market exposure, in particular short-term exposure to large firms.

Next, we estimate the demand for these management styles among Ethiopian firms. Using an obviously strategy-proof mechanism (Li, 2017), we elicit the preferences among senior human resource managers at relatively large Ethiopian firms for these types. We find that the authoritative type is strongly preferred by human resource managers, while preferences over the other three types are more heterogeneous. These types account for about half of the explainable variation in firms' preferences for these young professionals.

Finally, we show that the causal effect of a one-month management placement is driven by individuals from disadvantaged socioeconomic backgrounds, proxied by parental education. Among participants with no parent who completed primary school, control-group candidates are less likely to exhibit an authoritative management style and are assessed as worse by human resource managers than participants for whom at least one parent completed primary school; the placement eliminates these disparities. By contrast, for those with at least one parent who completed primary school, we detect no meaningful effect of the treatment.

1.3 Social Image, Organisational Values and Inclusion: Evidence from a Field Experiment

The third chapter studies a field experiment in Ethiopia focused on how institutions can align individual decision-making with broader institutional norms while preserving the expertise of decision-makers. The field experiment focuses on the behaviour of judges in a business plan competition in Ethiopia, randomly assigning judges to either receive a message on organisational values, share their decisions with peers to generate social image concerns, or both.

We find that male and female entrepreneurs performed equally well across control and treatment groups. However, the treatments increase the agreement among decision-makers both with each other and with external experts assessing the same candidates. Indeed, we find – using machine learning methods to study heterogeneity in treatment effects (Chernozhukov, Demirer, Duflo, and Fernández-Val, 2020) – that the treatments improve decision quality. Using a stylised theoretical model, we interpret this as the treatments reinforcing assessors' identities as representatives of the organisation rather than as individuals, increasing the weight they place on organisationally relevant information. Understanding the role of institutional design is key to developing policies within firms and at the policymaker level that align individual choices with organisational objectives and values.

1.4 Exploring Heterogeneity in Management: A Bayesian Approach to Identifying Complementarities

The final chapter studies heterogeneity in the combinations of management practices employed by firms. This chapter aims to contribute to the ongoing debate in the management literature about whether management should be seen as a technology that is either "good" or "bad" and transferable across contexts, or as something that should be designed to fit an organisation's specific environment (see, for example, Bloom, Sadun,

and Van Reenen, 2016a).

Specifically, I develop and estimate an ordinal mixture model that captures the combination of management practices employed by organisations. Using data from the 2006 World Management Survey on manufacturing firms, I find that while management practices are generally positively correlated across firms, clear differences exist in these correlations across various organisational structures and countries. Specifically, good practices related to monitoring and target setting are particularly strongly correlated. Additionally, the United States, United Kingdom, and Poland have particularly strong incentive practices relative to their other practices, conditional on the cross-country distribution of practices. I verify these findings using a model allowing for different latent types of firms to have different distributions in the combinations of management practices they employ. This novel empirical approach provides a framework to flexibly study heterogeneity in the combinations of management practices employed by firms.

Management as Rules: Evidence from an Adaptive Bayesian Questionnaire

Abstract

It is well understood that most organisational practices are communicated and understood through rules, but such rules have traditionally been difficult to measure well. This paper develops a novel Bayesian survey method to empirically measure the use of managerial rules within organisations. The method implements an adaptive survey, incorporating Bayesian optimisation, to first understand decision-making in a potentially high-dimensional space. It uses the information from this method to then directly ask about the rules driving patterns in these decisions. I apply this tool to measure the rules used by senior HR managers at Ethiopian firms. Descriptively, I find that this measure robustly and positively correlates with firm performance and helps explain variation in performance across firms, above and beyond more traditional measures of firm management. Using quasi-exogenous variation from firms' exposure to COVID-19, I further find that while managerial rules enhance firm performance, they also reduce resilience to economic shocks, suggesting a trade-off in the use of more rigid practices. The methods and findings in this paper motivate further research into the causal mechanisms underlying the effectiveness of managerial rules, and the factors influencing their heterogeneous adoption across organisations.

2.1 Introduction

Large and persistent productivity gaps across firms—even within narrowly defined industries—remain a central puzzle in economics (Syverson, 2011). While existing work shows that better management practices explain part of this variation (Bloom and Van Reenen, 2007; Gibbons and Henderson, 2012), we still know surprisingly little about the day-to-day decision rules that managers use and how these rules affect firm performance. Because many widely adopted 'practices' - from Lean to scientific management (Press, 1989; Taylor, 1911) - are themselves codified systems of rules, focusing on rules allows us to observe the micro-mechanism behind these aggregate practices. Economic theory also points to rules as key drivers of relational contracts (Alonso and Matouschek, 2008; Armstrong and Vickers, 2010), organisational resilience (Aghion, Bloom, Lucking, Sadun, and Van Reenen, 2021; Li et al., 2022), and firm-level productivity differences (Ellison and Holden, 2013).

Conceptually, I define an organisational rule as a heuristic decision procedure that is commonly known and consistently applied within the firm. Understanding how organisations use such rules is important not only for explaining variation in productivity, but also for analysing how firms manage trade-offs between consistency and flexibility. Rules can facilitate delegation, improve accountability, and reduce the cognitive costs relative to discretionary decision-making, especially in hierarchical or resource-constrained environments. At the same time, excessive reliance on rules may limit adaptation in turbulent or high-uncertainty settings. This trade-off, as shown empirically in Englmaier, Galdon-Sanchez, Gil, Kaiser, and Strandt (2020), is central to theories of organisational design.

However, despite their theoretical and empirical importance, organisational rules have thus received little empirical attention, largely because they are difficult to observe and measure. Rules are typically heuristics: simplified decision procedures based on a subset of available information. For example, a manager might, as a rule, promote employees every two years if their performance exceeds a certain threshold. I interpret such

heuristics as organisational rules when they are commonly known and consistently applied to similar decisions within a firm. Rules may be formal – documented in official guidelines – or informal, emerging through shared norms or repeated practice. Identifying these patterns requires both understanding how managers make decisions and distinguishing whether those decisions reflect personal judgement or codified procedures. Existing surveys of management practices do not capture this distinction, limiting our ability to assess how rules shape firm outcomes.

In this paper, I develop a novel tool to measure how firm managers use rules to make decisions. I apply this to a sample of senior HR managers at medium-sized and large Ethiopian firms. I then show that this measure of rules is positively and robustly correlated with firm performance even conditional on a wide range of firm observables. Finally, using quasi-exogenous variation, I show that this increased performance comes at a cost, with firms with more rule-based management being more negatively affected by COVID-19 restrictions implemented in Ethiopia.

In the first part of the paper I develop a tool to measure how managers use rules. The central idea is straightforward: if a manager relies on rules, her decisions should follow consistent patterns: specific facts leading to specific actions. But uncovering such patterns is challenging. There are many factors that might affect a decision, and too many possible combinations to ask about them all directly. The logic is similar to studying judicial reasoning: judges don't explain every possible case, but by observing how they decide a few well-chosen cases, we can learn the principles they apply. I use the same idea to study managers. I present them with a small number of carefully selected hypothetical decision problems and record their choices and the information they use to come to a decision. Each decision problem varies key facts – such as an employee's performance or tenure – to learn about patterns in a respondent's decisions. Crucially, I then confirm whether the observed patterns in decisions reflect personal judgement or rules commonly used within the organisation to ensure these are organisational rules rather than manager-specific heuristics. This approach allows me to measure rule-based behaviour without needing managers to describe those rules explicitly, while validating that patterns in

decisions are indeed driven by rules.

To choose the sequence of questions, I build on the literature in Bayesian optimisation to maximise the rate of learning about a respondent's decision making. This method uses a Gaussian process to model similarity between a respondent's decisions within a defined set of decision problems, assuming similar problems yield similar solutions. I then choose the question that, in expectation, is most informative about a respondent's decisions for all decision problems in this set. As I show in Section 2.7 this significantly increases the rate of learning about a respondent's decisions.

Turning to the empirical application, I implement this method in a new survey of 192 senior HR managers at medium-sized and large firms in Ethiopia. The sample spans a wide range of industries, including manufacturing, services, and finance, and captures variation in firm size, ownership, and management quality. Interviewing HR managers allows me to study relatively comparable decisions around hiring and compensation within this heterogeneous set of firms. Ethiopia provides a particularly relevant setting for studying organisational rules. As shown in Chapter 3 of this thesis and in Dahlstrand, László, Schweiger, Bandiera, Prat, and Sadun (2025), rule-based management appears to be especially valued in low-income countries, where firms often operate with a limited number of layers of middle management staff and face high costs of supervision and delegation (Bassi, Lee, Peter, Porzio, Sen, and Tugume, 2023). In such environments, clear and commonly understood rules may offer an essential mechanism for maintaining consistency, accountability, and control as firms grow.

Empirically, I begin by documenting substantial variation in the extent to which managers rely on organisational rules. Rule-based decision-making is strongly associated with observable firm characteristics: it is more prevalent in subsidiaries of larger firms, and correlates positively with the quality of HR practices and managers' beliefs about good management. I then show that rule-based management is predictive of firm performance. Firms with more formal rules tend to report higher revenue, profits, and profit margins. The predictive power of rule use exceeds that of standard management practice scores, such as the MOPS index, and remains robust across a range of specifications.

Having established the adoption of formal rules is positively correlated to performance, I test how the adoption of rules relates to firms' resilience to external shocks. Using data on the effects of COVID-19 restrictions, I find that firms with more rule-based management were more likely to experience disruptions in cash flow, postponed purchases, and reductions in profit during the pandemic. These findings are consistent with recent theoretical work suggesting that rigid rule structures can limit organisational adaptability in turbulent environments (Aghion et al., 2021; Li et al., 2022). Together, the results highlight both the benefits and costs of rule-based management: rules can support performance under normal conditions, but may impose constraints when flexibility is required.

This chapter makes three key contributions to the literature. First, I develop and apply a Bayesian adaptive questionnaire to feasibly measure organisational rulebooks used by managers. This contributes to the broad literature on unveiling the black box of management, including work on management practices (Bloom and Van Reenen, 2007), relational contracts (Gibbons and Henderson, 2012), CEO behaviour (Bandiera, Prat, Hansen, and Sadun, 2020), autonomy (Aghion et al., 2021), and delegation (Akcigit, Alp, and Peters, 2021a). The ability to measure organisational rules opens up a new research agenda in management and relational contracting. The developed methodology allows researchers to study questions related to better designing and targeting interventions to increase productivity growth, such as management interventions and providing access to capital.

Second, I contribute to the literature on management practices in firms in low-income countries. I show that implementing rules appears to be an important element of what managers do (Gibbons and Henderson, 2012), and that measuring rules can help explain heterogeneity in firm performance (Atkin, Khandelwal, and Osman, 2019; Syverson, 2011). This is important, as rules appear to be important in reducing the cost of delegation, which is key in low-income countries where this cost is typically very high (Akcigit, Alp, and Peters, 2021b; Bloom, Eifert, Mahajan, McKenzie, and Roberts, 2013a; Bloom, Genakos, Sadun, and Van Reenen, 2012a). This is in line with Bassi

et al. (2023)'s suggestion that reducing the cost of delegation and increasing managers' span of control are crucial to unlock economies of scale in sub-Saharan Africa. Finally, I document a trade-off between performance and resilience using a direct measure of rules used by managers, further developing the finding of Englmaier et al. (2020) that more rule-based management practices result in a similar trade-off – such practices perform well in stable periods but underperform during economic shocks.

Third, the Bayesian adaptive questionnaire I develop differs from existing methods by focusing on maximizing learning about a full set of related outcomes, rather than identifying specific parameters or finding a maximum. Existing methodology either focuses on finding a maximum (Frazier, 2018; Li, Raymond, and Bergman, 2020), on identifying specific parameters (Chapman, Snowberg, Wang, and Camerer, 2024), treatment effects (Caria, Gordon, Kasy, Quinn, Shami, and Teytelboym, 2024; Kasy and Sautmann, 2021), or on a one-dimensional function (Callander and Matouschek, 2019). Methodologically, the innovation in this paper is to use the principles from Bayesian optimisation to learn about a set of related decision problems. This method is more flexible than existing methods of optimal learning, but can match the performance of recent methods in adaptive choice experiments as I show in section 2.7.3.

The remainder of the paper proceeds as follows. Section 2.2 provides details on the Bayesian Adaptive Experimentation algorithm, and on how this algorithm is implemented to measure organisational rules. Section 2.3 describes the sample and the field implementation of the tool. Section 2.4 looks at how rules differ across firms and managers. Sections 2.5 and 2.6 explore the relationship between rules and respectively firm performance and resilience to economic shocks. Section 2.7 evaluates the performance of the adaptive sampling algorithm using both simulated and the collected data. Finally, Section 2.8 discusses the results.

2.2 Measuring Organisational Rules

To measure organisational rules empirically, I begin with a simple intuition: if a manager uses rules to make decisions, then their choices should follow patterns that reflect those

rules. For example, if a manager has a rule to never give a pay rise to an employee who has recently received one, this should be reflected in their decisions. I translate this intuition into a measurement strategy by first asking a respondent about their decisions in a set of hypothetical decision problems, and then confirming whether patterns in their decisions are driven by rules used in the firm or by individual preferences.

Building on this idea, I design a survey where managers respond to cases within hypothetical decision-making scenarios. A scenario consists of a set of related cases about, for example, a decision related to compensation. Each individual case presents a candidate with specific observable traits – for example, high performance, low pay (below median), and high mobility – and asks the manager whether they would grant a pay rise, and what information they are using to make this decision. These yes/no responses reveal how different attributes influence decisions. By also asking respondents which attributes they are actually using to make a decision (and only revealing the attributes that they want to use), I more directly elicit patterns in decisions.

The key methodological step is to reverse-engineer candidate rules from observed choices. A rule, in this context, is defined as a Boolean statement that links a combination of observable facts to a binary action. For instance:

> *"If an employee is low-paid, hard to replace, and high-performing, then grant a pay rise."*

Once such candidate rules are inferred from the observed patterns, I ask each manager to confirm whether these rules reflect formal or informal policies within their organisation, or whether they are based on personal discretion instead. Only rules confirmed as shared organisational practices are included in my measure of rule-based management.

To measure the extent to which managers use rules to make decisions, I introduce the concept of the breadth of the rulebook. Breadth is defined as the share of all possible combinations of attributes of a case for which a decision is governed by an organisational rule. So, for example, if a scenario has d binary characteristics, which can be combined 2^d possible ways, a rulebook has a breadth of 0 if it covers none of these combinations, a rulebook with a breadth of 0.5 covers $2^d/2$ combinations, and a rulebook with a breadth

of 1 covers all combinations. This measure thus abstracts away from the content of the rules, to instead focus on the extent to which managers use rules within the scenarios I develop.

In this methodology, a central challenge is that the number of possible scenarios grows exponentially with the number of characteristics: with d binary attributes, there are 2^d possible scenarios, and 2^{2^d} possible Boolean rules. Since it is infeasible to ask about every scenario, I employ an adaptive survey method that selects each new case to be maximally informative, based on the manager's previous answers. This adaptive mechanism uses a Gaussian Process model to approximate the manager's decision function and chooses each next case to minimise the expected posterior entropy. This means I choose a sequence of questions to, in expectation, reduce uncertainty about the underlying decision rule as efficiently as possible. Intuitively, the algorithm targets those cases about which it holds uncertain beliefs, to efficiently discriminate among competing rules.

The three steps of this methodology can be summarised as follows:

- **Adaptively select a sequence of scenarios:** Start with an initial seed scenario, then repeatedly *(i)* update a Gaussian Process surrogate of the manager's decision function with the latest yes/no response and *(ii)* pose the next case that minimises expected posterior entropy. Allow the respondent to choose which subset of attributes to observe to understand what information is used to make decisions. Repeat this process for a fixed number of eight questions.

- **Reverse engineer candidate rules:** From the accumulated pattern of choices, derive Boolean "if–then" statements that link combinations of observable attributes to the pay-rise decision. For example, *If low-paid & hard-to-replace & high-performing → grant a pay rise.* These statements constitute the set of candidate organisational rules.

- **Validate candidate rules:** Present each inferred rule back to the manager, asking whether it represents a formal, documented rule, an informal, commonly known rule, or personal discretion. Use only those inferred rules that respondents confirm

as shared organisational practices–formal or informal–as part of the final rulebook.

The methodology targets both implementation (what managers actually do) and perception (what they say they do), even though the data come from hypothetical scenarios. First, the adaptive scenario module reveals whether a manager's answers are consistent with any rules and identifies specific rules that fit those answers. Only after that trimming can we pose the direct question: "Is this pattern based on a formal rule, an informal norm, or mere discretion?". Including only the first step would not allow me to understand whether patterns in decisions are actually based on organisational rules. Including only the second step by directly asking about rules risks experimenter-demand effects, with managers overstating how rule-based their decision-making is. This latter concern is partially remedied by the requirement to at least make consistent decisions in the first stage to claim to be using rules.

The remainder of this section defines rules and why an adaptive algorithm is needed to measure these, before turning to the specific details and implementation of this methodology.

2.2.1 Defining rules

In this section, I formally define decision rules as Boolean functions that map observed attributes of a decision problem to a decision. A distinct characteristic of these rules is that they rely on a subset of the available attributes, creating a sufficient condition for making a decision without (necessarily) using all available information. I then define the breadth of a rule as the proportion of decision problems (i.e., combinations of attributes) it applies to, so a rule that ignores many characteristics is considered broad. Finally, I define a rulebook as a collection of rules used by a decision maker, and the breadth of a rulebook as the proportion of observed attribute combinations covered by any rule within it.

I define a scenario as a set of combinations of d binary characteristics. Let each

combination of these characteristics:

$$x \;=\; (x_1, \ldots, x_d) \in \{0,1\}^d,$$

denote a "case". A decision is $y \in \{0,1\}$.

A decision rule is then a mapping that translates a subset of the binary characteristics of a case to a decision, specifically:

$$h : \mathcal{D}_h \subseteq \{0,1\}^d \;\longrightarrow\; \{0,1\}.$$

The rule applies on the domain $\mathcal{D}_h$, which collects exactly those states for which the rule is meant to apply. I say the rule *uses* the subset of attributes $S_h \subseteq \{1, \ldots, d\}$ if $h(x) = h(x')$ whenever $x_i = x'_i$ for all $i \in S_h$. This means that if two cases are governed by the same rule, they yield the same decision regardless of characteristics outside of S_h.

A key feature I will focus on empirically is the breadth of a rule. Intuitively, this is the share of the decisions in this scenario space that is governed by a single rule - so a rule that applies for every combination of these d binary characteristics has a breadth of 1, and a rule that applies to a single case has a breadth of $\frac{1}{2^d}$. Specifically, the breadth of a rule is defined as:

$$B(h) \;=\; \frac{|\mathcal{D}_h|}{2^d} \qquad \in [0,1].$$

A rule that ignores many attributes (small $|S_h|$) has a large domain and therefore high breadth.

Finally, I am going to call the combination of rules used by a respondent a rulebook. So, I let $\mathcal{R} = \{h_1, \ldots, h_k\}$ be the set of rules the manager says they follow. I am then going to focus, again, on the share of the decisions of a manager that is governed by this rulebook - the breadth of the rulebook, or equivalently the domain of the rulebook $\mathcal{D}_\mathcal{R}$: $\mathcal{D}_\mathcal{R} = \bigcup_{j=1}^{k} \mathcal{D}_{h_j}$, and

$$B(\mathcal{R}) \;=\; \frac{|\mathcal{D}_\mathcal{R}|}{2^d},$$

the breadth is the fraction of all possible cases for which the manager can point to a rule

to make a decision.

2.2.2 Measuring rules based on individual decisions

Here, I expand on the intuition that if managers make decisions based on rules, this would manifest itself as patterns in their decisions. Specifically, if a manager implements a rulebook $\mathcal{R}$ all decisions for cases $c \in \mathcal{D}_{\mathcal{R}}$ would be made based on a rule. I can then ask the manager to decide on every possible combination of characteristics, and use this information to generate a set of decision rules that would justify these patterns. Finally, I can then ask whether these decision rules are driven by actual rules within the firm.

To illustrate, throughout this section I use a simple three-attribute example. In this example, the question is whether to give an employee a pay rise as a function of a set of binary characteristics. Each hypothetical employee is described by a binary vector $(x_1, x_2, x_3) \in \{0, 1\}^3$, where $x_1 = 1$ indicates high performance, $x_2 = 1$ indicates above average pay, and $x_3 = 1$ indicates high mobility (the worker could easily leave the firm). The manager's decision is $y \in \{0, 1\}$ with $y = 1$ meaning "grant a raise."

Suppose we wanted to elicit a manager's decisions, and the rules used to make those decisions in this scenario. In this scenario, there are eight possible cases and we could ask the manager exhaustively what she would do in each of these scenarios. Suppose that, after asking about eight cases, we see that the respondent always gives a pay rise when performance is high, and never gives a pay rise when performance is low. This is consistent with a few possible organisational rules, two obvious examples are (i) that the respondent always gives a pay rise when performance is high, and (ii) that the respondent never gives a pay rise when performance is low. To identify whether these are actual rules in the firm, or whether these patterns are just based on the discretionary choices of the managers, I then validate these rules with the respondent. Specifically, I ask the respondent whether the firm has a rule to give a pay rise whenever $x_1 = 1$, and whether she has a rule to never give a pay rise whenever $x_1 = 0$ or whether she has discretion in this case.

The issue with this method however is that the number of combinations of characteristics, and the number of potential rules, increases exponentially with the number

of characteristics. With, for example, $d = 7$ characteristics there are $2^7 = 128$ possible combinations of these characteristics. This motivates choosing a sequence of cases that is particularly informative about a respondent's decision-making, and allowing the respondent to choose which subset of attributes of a case to observe.

2.2.3 Adaptively selecting a sequence of scenarios

To sample a set of cases within each scenario to maximise the rate of learning about a respondent's decision-making, I develop a novel method of choosing a sequence of questions, a "Bayesian adaptive questionnaire". This allows me to generalise the approach described for $d = 2$ to a more high-dimensional setting. Intuitively, the approach leverages the idea that similar questions elicit similar answers from the same respondent, and that the way attributes affect decisions shows some similarity across respondents. Consequently, once we observe a respondent's previous answers, certain questions become more (or less) informative about their decision-making patterns. This approach improves the efficiency of information gathering within a fixed amount of time, which is valuable in settings where asking a large number of questions is either expensive or increases noise.

Consider that each scenario consists of a set of cases with d characteristics x and binary outcomes y. I denote the full set of cases by $C = \{X, Y\}$. I partition this set C into a set of observed cases, $C_o = \{X_o, Y_o\}$, for which I have collected responses, and a set of unobserved cases, $C_u = \{X_u, Y_u\}$, for which I have not yet collected responses. As before, a single case is denoted by $c = \{x, y\}$.

Each scenario thus consists of 2^d total cases, and my objective is to best predict the respondent's decision in each case by observing a limited number of the binary outcomes in Y. The outcomes Y are related to the characteristics X through an unknown data-generating function $Y = g(X)$. Instead of selecting all cases in C_o at once, I adopt a sequential, or "greedy", approach for computational efficiency, inspired by the Bayesian optimisation literature (Frazier, 2018; Jedynak, Frazier, and Sznitman, 2012). This involves selecting and presenting questions one at a time, and updating the model's beliefs after each response.

The model consists of two main components - the first standard and the latter more novel. The first is a surrogate function $\hat{Y} = f(X, Y_o)$ to approximate the true function $Y = g(X)$. This function leverages both prior information and the observed outcomes Y_o to form predictions for the full set of cases. The second component is an objective function that quantifies the expected reduction in uncertainty from asking each potential question. By evaluating this objective function for each potential next question, I can select the next question that maximises the expected rate of learning. The remainder of this section describes these components.

I pick the next case that in expectation reduces predictive entropy over the entire 2^d space as much as possible. Specifically, the objective function for the algorithm is to minimise the expected posterior entropy over the full set of cases in C, by choosing a case x for question t such that:

$$x_t^* = \arg\min_{x \in X} \mathbb{E}_{y_x} \left[\sum_{x' \in X} H(y_{x'} \mid C_o) \right], \tag{2.1}$$

where

$$H(y_x \mid C_o) = -\sum_{y \in \{0,1\}} P(y_x \mid C_o) \log P(y_x \mid C_o).$$

and y_x is the decision y in case x. The expectation is over the prior predicted probabilities that $y_x = 1$ and $y_x = 0$. This ensures that each selected case, in expectation, minimises uncertainty about how the respondent will decide in the full set of cases, conditional on the set of observed cases at that point in time.

Next, for each respondent, some unknown data-generating process $Y = g(X)$ maps observable facts to decisions. As this function is unknown, I need a flexible function to approximate this that allows me to quantify uncertainty, generalise from sparse data and can be updated with incremental data. I follow the literature on Bayesian optimisation (Frazier, 2018), and approximate this function using a Gaussian Process (GP) regression model as my surrogate function $f(X_o, Y_o)$. The GP approximates the mapping from attribute vectors to binary decisions, not the underlying logic or rules themselves.

A Gaussian Process provides a flexible, non-parametric method to model complex

relationships by placing a joint Gaussian distribution over the function values corresponding to different inputs. A Gaussian Process is specified by a mean function $m(X)$ and a covariance function (kernel) $k(X, X')$, and is denoted as:

$$f(X) \sim \mathcal{GP}(m(X), k(X, X')).$$

This allows me to incorporate prior beliefs and observed data to predict outcomes for unobserved cases. The choice of mean function and kernel can be informed by theoretical considerations, prior data, or machine learning techniques, and they determine the GP's ability to approximate the true function $g(X)$.

With a GP prior and noise-free observations, the joint distribution of the function values at the observed inputs X_o and the unobserved inputs X_u is given by:

$$\begin{bmatrix} f(X_u) \\ f(X_o) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(X_u) \\ m(X_o) \end{bmatrix}, \begin{bmatrix} K(X_u, X_u) & K(X_u, X_o) \\ K(X_o, X_u) & K(X_o, X_o) \end{bmatrix} \right),$$

where $K(X_i, X_j)$ denotes the kernel evaluated at the inputs X_i and X_j. Using the properties of multivariate normal distributions, I can derive closed-form expressions for the posterior mean ($\mathrm{E}(Y|Y_o)$) and covariance ($\mathrm{var}(Y|Y_o)$) of the full set of outcomes Y, conditioned on the observed data (X_o, Y_o).[1]

Since my outcomes Y are discrete (i.e. binary decisions) rather than continuous, I extend the GP regression framework to classification tasks. I use the latent function $f(X)$ modelled by the GP and apply a logistic function to map it to the probability of each outcome:

$$P(y = 1|X) = \sigma(f(X)) = \frac{1}{1 + e^{-f(X)}},$$

This allows me to compute the probability distribution over possible outcomes for each case.[2]

[1] This specification gives me significant flexibility in how to specify the mean function and kernel. The performance improves the closer the mean function and kernel approximate the true data-generating process. These functions can be modelled using a theoretical framework (e.g., when measuring preferences as per Chapman et al. (2024)), machine learning algorithms, or any other function of the data that generates a valid covariance function.

[2] In practice I use the Laplace approximation, see e.g., (Rasmussen and Williams, 2005)

Combining this Gaussian Process Classification model with the objective defined in equation 2.1 allows me to choose an optimal question at each step. By repeating this process–tselecting the most informative question, obtaining the respondent's answer, and updating the surrogate function–I iteratively refine my predictions for the full set of cases C. This adaptive approach ensures that I gather the most valuable information within the limited number of questions I can ask.

2.2.4 An illustrative example of adaptive sampling

Now, let us return to the previous example of the pay rise scenario with three binary features. Suppose the respondent has already answered two cases: $(1, 1, 0) \mapsto 1$ and $(0, 0, 0) \mapsto 0$. The Gaussian Process (GP) model uses these to update its beliefs over the decision function $f(x)$. Crucially, the GP generalises from these responses to all nearby cases: for example, it predicts that $(1, 1, 1)$ is likely to also result in $y = 1$ because a more mobile employee is more likely to get a pay rise all else equal.

The adaptive algorithm then seeks the next scenario x that most reduces uncertainty about h^* across all possible cases as defined in equation 2.1. As an example, the case $(1, 1, 1)$ is likely to be quite uninformative as it is predictable (if $(1, 1, 0)$ yields $y = 1$, so should this case). Instead, the case $(1, 0, 0)$ lies between the two previously observed cases, making observing this case more informative. There is uncertainty about this case as two cases with a different decision y are similar to it, and it is comparable to a set of other cases with relatively high uncertainty. This is thus a good candidate to ask about relative to $(1, 1, 1)$. This example illustrates that adaptive sampling in a GP framework exploits both local uncertainty and spatial correlation: each answer informs the model about the broader structure of the decision rule h^*, as proxied by patterns in a respondent's decisions.

2.2.5 Allowing respondents to choose the attributes they observe

An additional feature of this methodology is that respondents are not required to observe all scenario attributes before making a decision. Instead, they are explicitly asked to choose which attributes they would like to see for each hypothetical case. This design serves two purposes. First, it mirrors real-world decision-making processes, where managers often prioritise a subset of information when making a choice. Second, it provides additional data on which attributes are considered relevant by the respondent in the context of each decision.

Formally, for each scenario $x \in \{0,1\}^d$, let $x^{\text{obs}} \subseteq x$ denote the subset of attributes the respondent chooses to observe before responding. The observed decision y is then interpreted as a function of only x^{obs}. This creates a partial information setting, where the GP model conditions only on the observed inputs x^{obs} when updating its posterior beliefs.

To reduce cognitive burden, respondents may specify a default subset of attributes that they wish to always observe. These defaults are presented automatically in each scenario, with the option to request additional information as needed. This structure ensures that the algorithm not only learns the mapping from scenarios to decisions, but also identifies which features are actively used in making those decisions.

2.2.6 Mapping observed decisions to candidate rules

To identify candidate rules from a respondent's answers, we first partition the full set of possible decision scenarios into three categories: Those for which the respondent says yes ($y = 1$), those for which the respondent says no ($y = 0$) and those that are not covered by the respondent's decisions, or for which contradictory decisions were made. This method is based on the assumption that a respondent will not give answers that violate an organisational rule.

A *candidate rule* is any Boolean "if–then" statement that uses a subset of scenario

attributes to correctly predict the manager's decision in a way that is perfectly consistent with either the "yes" and "no" sets. This means that for any case covered by a rule that yields $y = 1$, we have observed the manager responding with $y = 1$ in the previous stage of the data collection.

Suppose, as before each decision scenario includes three binary attributes: performance (x_1), pay level (x_2), and mobility (x_3). A manager's decisions may be consistent with the following mapping:

- Grant a pay rise for employees who are high-performing, low-paid, and mobile: $(x_1 = 1, x_2 = 1, x_3 = 1) \rightarrow y = 1$,
- Deny a pay rise for all other combinations: $y = 0$; so effectively when either (a) the employee is not high-performing, (b) the employee is already highly paid or (c) the employee has limited outside options.

  From this, one candidate rule is:

  "If an employee is high-performing, low-paid, and mobile, then grant a pay rise."

This rule is valid because it perfectly aligns with the manager's observed responses.

Another candidate rule is:

"If an employee is not high-performing, do not grant a pay rise."

This method thus generates a large number of candidate rules, including all narrower rules nested within these two broad rules.

2.2.7 Validating candidate rules

We now have a large set of candidate rules to validate. To decide which rules to show, I order candidate rules from broadest to narrowest. For example, a rule like "If an employee does not perform well, do not give a pay rise" is very broad as it ignores any other facts. Instead, a rule like "If an employee does not perform well, but is poorly paid and can

easily leave, give a pay rise" is narrower as it incorporates more information to come to a decision – and thus applies to fewer cases.

I then present up to the ten broadest candidate rules in turn to the manager. For each rule the manager indicates (a) whether it is used in her firm (either explicitly as stated here, or as part of some broader rule within the firm) and, if so, (b) whether it is formal (codified) or informal (shared but unwritten). Once a broad rule is confirmed, any narrower rules nested within it are skipped as they would not add to the breadth of the rulebook. Only candidate rules explicitly affirmed in this step are classified as rules. The output is a validated set of formal and informal rules that is directly comparable across managers and firms.

I intentionally cap the number of candidate rules proposed at ten to avoid fatiguing respondents. This restriction influences how the algorithm infers the breadth of a rulebook. When a manager relies on several narrow rules that always lead to the same action, those rules effectively merge into one broader rule in the algorithm's eyes, because identical consequences across cases are interpreted as a single "if–then" statement. The estimated breadth therefore captures the breadth of the rulebook, but not the number of clauses it contains. A different outcome arises when a manager applies many narrow rules that give different actions to very similar cases. With only ten observations, the algorithm then produces ten very narrow rules, and a rulebook with a narrow breadth due to the limit in the number of questions. This is in some sense a flaw in the methodology, but consistent with the way I conceptualise rules as stable, commonly understood guidelines rather than case-by-case protocols.

2.3 Management as rules in Ethiopian firms

I implement this methodology to measure the use of rules by HR managers at Ethiopian firms. To do so I use two scenarios, one related to compensation and one to hiring. The key question for the respondent is whether to grant a pay rise or hire a candidate, respectively, as a function of seven binary characteristics. This section first details the scenarios and provides a further example, before turning to the sample used in the survey

and the details of the implementation.

2.3.1 The Human Resource Scenarios

Box 2.1 and Box 2.2 detail the pay rise and hiring scenarios respectively. In these scenarios, a colleague – role-played by the enumerator – asks for the respondent's advice. This approach creates a realistic context in which the respondent is aware they cannot consider all of the employee's characteristics when making a decision. In both scenarios, we frame the discussion by asking the respondents to identify the most common position in their firm, and to answer the questions as if relating to this position.

The pay rise scenario (Box 2.1) is presented as follows: "Suppose I, as your colleague, tell you an employee is unhappy with their pay and would like a rise to the next point on the pay scale, and I ask for your decision. Before making a decision, you can ask to learn about any of the following characteristics of the employee." The enumerator then reads out the seven characteristics in a random order, and shares a document with the respondent listing the full set of scenario characteristics for reference. The specific characteristics, for both scenarios, were selected based on piloting the module in Ethiopia.

This box also provides an example of how the survey may play out in practice. The column "example case" provides a draw of seven binary characteristics that may be generated by the adaptive algorithm, and given to the respondent. This example case fixed the seven realisations of the individual characteristics. The respondent can then, iteratively, choose which of these individual characteristics to observe. For example, as indicated in the observed column the respondent may wish to ask about performance, mobility and pay, observe the fixed characteristics for this case, and then make a decision not to give a pay rise. If this were the only question asked, this would yield the candidate rule: "If an employee has above average performance but has above average pay and cannot easily leave for another firm, do not give them a pay rise".

The hiring scenario (Box 2.2) is presented as follows: "Suppose I, as your colleague, tell you I have a candidate for an entry-level position at your firm for which you are hiring. I would like your decision on whether you would hire this candidate. Before

Table 2.1: The pay rise scenario

"Suppose I, as your colleague, tell you an employee is unhappy with their pay and would like a raise to the next point on the pay scale and I want your decision. Before making a decision, you can ask to learn about any of the following characteristics of the employee."

Abbreviation	Characteristic	Options	Example Case	Observed
Performance	The employee's performance compared to other employees with similar jobs.	Above/below average	Above average	✓
Likes	Your colleague personally likes the employee.	Yes/No	Yes	
Mobility	The employee could easily leave for another firm.	Yes/No	No	✓
Pay Rise	The employee has had a pay rise in the past year.	Yes/No	Yes	
Pay	The employee's pay is above average at the firm.	Yes/No	Yes	✓
Tenure	The employee has worked at the firm for more than two years.	Yes/No	Yes	
Gender	The employee's gender.	Male/Female	Male	

Notes The pay rise scenario asks the respondent about compensation. The scenario has seven specific characteristics, and the respondent knows what these are. Each characteristic can take two values. The final column depicts an example case that could be selected by the adaptive algorithm.

Table 2.2: The hiring scenario

"Suppose I, as your colleague, tell you I have a candidate for an entry-level position at your firm for which you are hiring. I would like your decision on whether you would hire this candidate. Before making a decision, you can ask to learn any of the following characteristics:"

Abbreviation	Characteristic	Options
Enthusiasm	The candidate showed enthusiasm for the job.	Above/below average
Interview	The candidate performed well in the interview.	Yes/No
English	The candidate speaks English.	Yes/No
Culture	The candidate fits with your firm's culture.	Yes/No
Experience	The candidate has relevant experience at another firm.	Yes/No
University	The candidate has a university degree.	Yes/No
Gender	The candidate's gender.	Male/Female

Notes The hiring scenario asks the respondent about hiring a new employee. The scenario has seven specific characteristics, and the respondent knows what these are. Each characteristic can take two values listed in the "Options" column.

making a decision, you can ask to learn about any of the following characteristics." The enumerator then again shares the list of characteristics and reads these out.

During the survey, the Bayesian adaptive questionnaire selects a sequence of eight cases (a combination of the seven characteristics) to present to the respondent for each vignette. For each case, I observe the respondent's decision and the information they used to make that decision. The algorithm interprets the respondent's decision as indicative of their choice in all cases sharing the observed characteristics. For instance, if the respondent decides to grant a pay rise to an employee with above-average performance and pay but below-average mobility, the algorithm assumes this decision applies to all cases with these characteristics. These cases are added to the set C_o and are used to update the model's beliefs. Based on the accumulated information from previous cases, the algorithm selects the next case to present.

The sequence of the respondent's decisions and the information used to make these decisions allows the full set of cases to be divided into three groups: the set of cases for which the respondent has indicated they will not give a pay rise, the set of cases for which the respondent has indicated they will give a pay rise, and the remaining cases for which the respondent has not shared a decision. Based on these three sets of cases, I generate a set of hypothetical rules, i.e., mappings from a subset of characteristics to decisions, that are consistent with the respondent's decisions. For example, if an employee performs well, is currently paid relatively poorly, and could easily leave, they should receive a pay rise.

These hypothetical rules form the input for the second part of the questionnaire. In this second part, I directly ask respondents whether these hypothetical rules represent actual rules within their organisation. Starting with the broadest rule, i.e., covering the most cases, I ask the respondent whether it applies. If it does, I next inquire whether the rule is formal (documented and codified) or informal (generally understood but not documented). This process continues with less broad rules as necessary, allowing respondents to classify each of these hypothetical rules.

2.3.2 The sample

I apply the developed methodology to examine personnel rules in a sample of relatively large Ethiopian firms. These firms are broadly representative of businesses in Addis Ababa. Focusing on human resource management–a universal aspect of organisational practices–facilitates the comparison of the diverse set of firms in this sample.

The participating firms are medium-sized to large establishments, operating across various sectors of the economy. The median number of employees is fifty (25th percentile = 20, 75th percentile = 142). Most of the firms are located in Addis Ababa, while around 15% are situated in the mid-sized towns of Adama (80 km from Addis Ababa) and Bishoftu (40 km from Addis Ababa). The firms operate in a range of sectors, primarily manufacturing (40%), services (30%), and retail and trade (20%), with the remainder in other industries. Of the firms sampled, 20% are government-owned, while 80% are

privately owned. The sample is drawn from firms previously studied in Abebe, Fafchamps, Koelle, and Quinn (2019), with some additional, comparable firms added to the sampling frame.

The respondents are highly experienced managers responsible for HR decisions within their firms. Around 75 of the surveyed managers are either owners or managing directors, while the rest are senior managers with HR responsibilities. The managers have a median tenure of seven years at their current firms (mean tenure of nine years) and a median of fifteen years of total work experience (mean of eighteen years). Approximately 90% of the managers hold at least a bachelor's degree, and 65% have formal management education. The sample is restricted to managers who have the authority to make decisions related to hiring and compensation within their firm.

2.3.3 Implementation

A team of experienced enumerators conducted the survey. To ensure clarity and comprehension, the survey focuses on two scenarios, each with seven fixed binary characteristics, resulting in 128 (2^7) possible combinations. For each scenario, respondents were presented with eight distinct cases, each consisting of a unique combination of the seven characteristics. These cases were selected using the adaptive algorithm. The hiring and pay rise scenarios were detailed earlier in Section 2.3.

To implement the survey, I developed a custom application to be used by the enumerators in Addis Ababa. Since no reliable internet connection was available, all computations had to be performed locally on the enumerators' computers during the interviews. The application allowed enumerators to record respondents' answers, and then used the Bayesian adaptive algorithm to solve the optimisation problem of selecting the next question. Because the algorithm considers the full history of each respondent's answers – including the information they chose to observe – the sequence of questions was selected in real time. Following this, the survey application generated a set of heuristics consistent with the data collected on decisions. Enumerators then used the application to ask whether these heuristics represented organisational rules. Details of the questionnaire

are provided in Appendix 2.9.B.

At the start of each interview, respondents were asked to describe the duties of an employee in the firm's primary area of operation and to answer subsequent questions with this position in mind. This approach helped frame the discussion and allowed for control over the specific position discussed during the analysis.

Training data for the survey was collected from respondents within the same population by asking about eight cases selected randomly from the full set of potential combinations in the two hypothetical scenarios. These questions were identical to those in the main survey, except that respondents were shown all available information rather than choosing which characteristics to observe. This training data was used to fit the parameters of the Gaussian Process model employed in the survey.

The Gaussian Process model employed a linear mean function and a squared exponential kernel with automatic relevance determination. This configuration allowed the impact of each of the seven characteristics on the covariance between two cases to differ. A logistic link function was used to map the Gaussian Process outputs into probability space. Specifically:

$$m(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta} + \alpha$$

and,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{7}\frac{(x_{i,d} - x_{j,d})^2}{\ell_d^2}\right)$$

2.3.4 Identified patterns in firms' decisions and rules

Turning to the data collected from the adaptive Bayesian questionnaire, I first assess which characteristics respondents chose to observe and how these characteristics were used to define rulebooks. I define a rulebook as the collection of rules identified for a firm.

On average, respondents asked to observe four characteristics per decision (median = 4; 25th percentile = 3, 75th percentile = 5) in both scenarios. The identified rules tended to be broad and typically relied on only two characteristics (median = 2; 25th

percentile = 1, 75th percentile = 3). This pattern held across both scenarios. Firms used, on average, two rules per scenario (median = 2; 25th percentile = 1, 75th percentile = 4).

Table 2.3 summarises the characteristics observed and used in rules for the pay rise scenario. The most frequently observed characteristics were performance and tenure, while gender was the least observed. A similar trend was found in the probability of each characteristic being included in a rulebook. Interestingly, while performance was often used in decision-making, it featured less frequently in the derived rules. The last two rows in the table consider the probability that each characteristic is included in a rule that leads to a particular decision, as well as the average value of the characteristic when it is part of a rule leading to that decision.[3]

Table 2.4 presents a summary of the characteristics observed and included in rules for the hiring scenario. In this scenario, a wider range of characteristics were frequently observed, including enthusiasm for the position, interview performance, cultural fit, and experience. Gender was observed more frequently in the hiring scenario compared to the pay rise scenario; however, it was not much more likely to feature in a rulebook.

2.4 The relationship between rules, firms, managers and management

In this section, I examine how rules relate to firm and manager characteristics. I first consider how the use of rules correlates with the characteristics of the firms and managers. Then, I investigate whether firms with broader rulebooks implement different management practices. To measure the extent to which managers' decisions are constrained by rules, I focus on how comprehensive the rulebook is. Each scenario contains $2^7 = 128$ possible cases. I define the *breadth* of the rulebook as the proportion of cases in which the manager makes their decision using a rule—that is, the number of cases covered by rules divided by the total number of cases in the scenario. A breadth close to one indicates highly rule-based decision-making, whereas a breadth close to zero suggests mainly dis-

[3] Gender is encoded as 0 = female, 1 = male.

Table 2.3: Characteristics used in decision making in the pay rise scenario

	Performance	Like	Mobility	Pay rise	Pay	Tenure	Gender
	Probability a respondent asks about a characteristic in a specific case						
Observed by respondent	77.4%	50.0%	41.9%	40.1%	36.6%	68.2%	19.9%
	Probability that a characteristic features in a respondents' rulebook.						
Features in a rulebook	48.7%	42.3%	31.4%	21.2%	19.2%	53.8%	17.9%
Decision	Probability of a characteristic featuring in a rule making a specific decision.						
No Raise	31.2%	33.0%	32.1%	22.3%	20.5%	41.1%	18.8%
Raise	26.6%	25.7%	19.1%	11.8%	10.9%	35.2%	12.2%
	Probability of a characteristic being 1 conditional on being in a rule.						
No Raise	23%	46%	14%	48%	52%	26%	48%
Raise	93%	91%	81%	75%	52%	83%	62%

Notes This table depicts what information managers use in decisions and rules in scenario 1. For the first row, this depicts that probability that a characteristic is observed by a respondent when making a decision. The second row, this depicts the probability that, for each respondent, the characteristic features in at least one rule. The third rows depicts the probability each characteristic features in an individual rule that makes the decision to give the pay rise, and in a rule that makes the decision not to give a pay rise. The final set of results gives the average value of a characteristic, conditional on it featuring in a rule making a specific decision. The characteristics included are performance (1=high), like (1=colleagues likes the respondent), mobility (1 = could leave), pay rise (1 = had one in the last year), pay (1 = above average), tenure (1 = more than two years) and gender (1=male).

Table 2.4: Characteristics used in decision making in the hiring scenario

	Enthusiasm	Interview	English	Culture	Experience	University	Gender
	Probability a respondent asks about a characteristic in a specific case						
Observed by respondent	67.9%	66.8%	29.4%	67.6%	73.2%	49.6%	26.1%
	Probability that a characteristic features in a respondents' rulebook.						
Features in a rulebook	47.9%	55.2%	17.0%	47.9%	48.5%	46.1%	18.8%
Decision	Probability of a characteristic featuring in a rule making a specific decision.						
Don't hire	30.3%	27.6%	7.2%	34.9%	27.6%	23.7%	10.5%
Hire	26.2%	29.2%	7.6%	25.1%	27.8%	28.4%	10.3%
	Probability of a characteristic being 1 conditional on being in a rule.						
Don't hire	17%	19%	36%	17%	14%	42%	38%
Hire	72%	77%	43%	65%	74%	77%	61%

Notes This table depicts what information managers use in decisions and rules in scenario 2. For the first row, this depicts the probability that a characteristic is observed by a respondent when making a decision. The second row, this depicts the probability that, for each respondent, the characteristic features in at least one rule. The third rows depicts the probability each characteristic features in an individual rule that makes the decision to hire the candidate, and in a rule that makes the decision not to hire the candidate. The final set of results gives the average value of a characteristic, conditional on it featuring in a rule making a specific decision. The included characteristics are enthusiasm (1=showed enthusiasm), interview (1=performed well), English (1=speaks English), culture (1=fits with culture), experience (1=has relevant experience), university (1=has a university degree), gender (1=male).

cretionary decision-making. I calculate the breadth for the entire rulebook and separately for the set of formal rules, defining the *breadth* and the *formal breadth* of the rulebook, respectively.

2.4.1 Rules and firm observables

In this section, I conduct descriptive regression analyses to explore how firm characteristics relate to the breadth of the rulebook. The firm characteristics considered include firm size, sector, the type of employee position, and whether the firm is a subsidiary of a larger firm. Since each firm is observed in two scenarios, I obtain two measures of breadth per firm. To account for this, I run all regressions with standard errors clustered at the firm level. Specifically, for firms $f \in \{1, 2, .., F\}$ and $s \in \{1, 2\}$ for outcomes y and covariates X

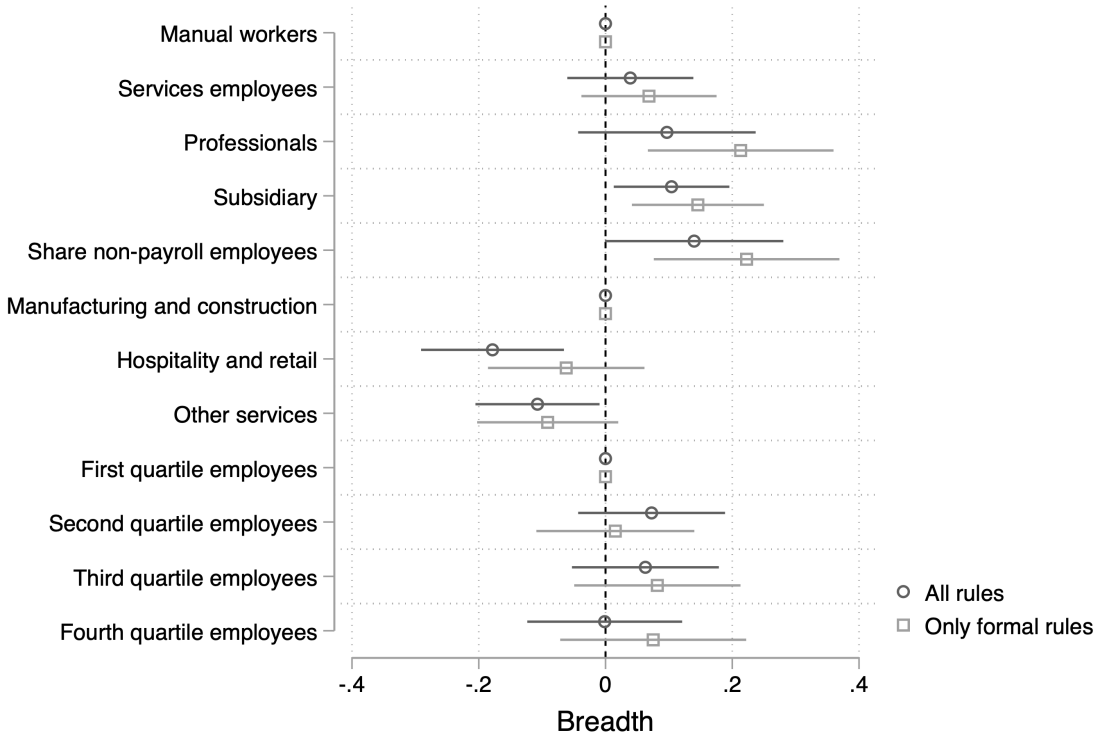
$$y_{fs} = \alpha + \boldsymbol{\beta} \boldsymbol{X}_f + \varepsilon_{fs}, \tag{2.2}$$

Figure 2.1 presents the results from these regressions. This figure shows three variables are correlated with the breadth of the rulebook used by firms: whether the main position being interviewed for is a professional position, whether the firm is a subsidiary of a larger firm, and the share of non-payroll employees at the firm. This relationship appears to be driven by formal rather than informal rules. These results suggest managers at different types of firms differ in the use of formal rules. Figure 2.1 presents a multivariate regression, Appendix Figure 2.6 shows the results are comparable in a univariate framework.

2.4.2 Which managers use rules?

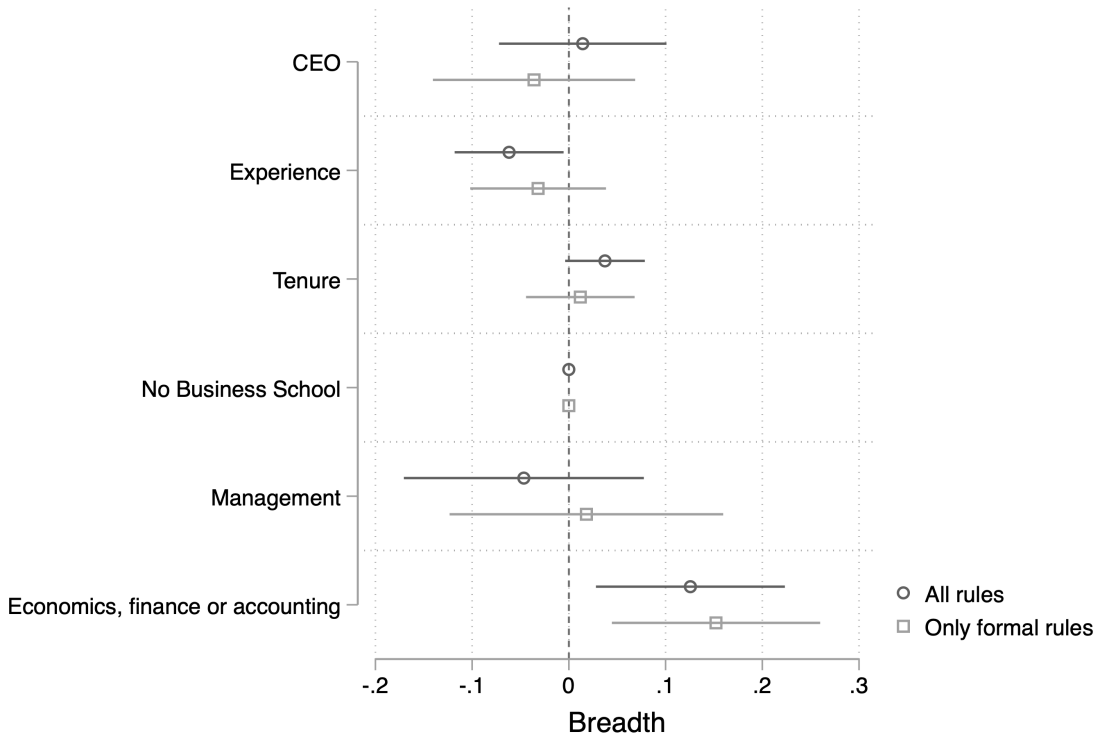
The following section focuses on whether different managers adopt different rules. Figure 2.2 summarises the results of regressions of the breadth of the rulebook on various manager characteristics. The results indicate that CEOs tend to use a narrower set of formal rules. Managers who studied economics, finance, or accounting at a business school tend to use broader rules, whereas managers who studied management at a business school

Figure 2.1: The relationship between firm observables and the breadth of the rulebook.



Notes This figure reports regression results from multivariate regressions of the breadth of the full and formal rulebook on various firm characteristics. These include employee type, categorised into manual workers, service employees, and professionals, with manual workers as the base category; the share of non-payroll employees; a subsidiary dummy variable where 1 denotes that the plant where the interview took place is a subsidiary of a larger entity; industry classification, which includes manufacturing and construction, retail and trade, and other service sectors such as tourism and education, with manufacturing as the base category; and firm size in terms of employment, included by quartile. Error bars represent 95% confidence intervals, with standard errors clustered at the firm level.

Figure 2.2: The relationship between manager characteristics and the breadth of the rulebook.



Notes This figure reports regression results from multivariate regressions of the breadth of the full and formal rulebook on various respondent characteristics. The CEO dummy indicates whether the respondent is the CEO or equivalent. Experience and tenure are respectively years of experience and years of experience at the firm, measured in units of five years. The management dummy indicates whether the respondent studied management at a business school. The economics dummy indicates if the manager studied economics, finance, or accounting at a business school. Error bars represent 95% confidence intervals.

do not differ significantly from those who did not attend a business school in their use of rules. Appendix Table 2.7 shows that the results are comparable in a univariate regression framework.

In the survey, I also elicit both a measure of how much the respondent trusts people in general and a measure of how much the respondent trusts employees as a manager. I find that managers who are more trusting in general implement broader rulebooks, but interestingly this effect only holds for respondents who do not also trust their employees. These effects are large and statistically significant, especially concerning the implementation of formal rulebooks. This relationship might reflect that managers who trust their employees, due to a well-functioning relational contract within the firm, may not feel the

need to implement extensive rules to manage them. This finding aligns with Bloom et al. (2012a), who find that higher trust within multinational organizations results in greater degrees of decentralisation. Similarly, I observe a less rule-based environment when there is more trust in employees. Table 2.5 summarises these results.

Table 2.5: The relationship between organisational rules and trust

	(1) Rule Breadth	(2) Formal Rule Breadth	(3) Rule Breadth	(4) Formal Rule Breadth
General Trust	0.235***	0.434***	0.206***	0.338***
	(0.05)	(0.07)	(0.06)	(0.07)
Employee Trust	-0.139*	0.034	-0.127	-0.005
	(0.08)	(0.09)	(0.08)	(0.09)
General Trust X	-0.192*	-0.418***	-0.160*	-0.311***
Employee Trust	(0.10)	(0.11)	(0.10)	(0.11)
Constant	0.618***	0.306***	0.611***	0.227***
	(0.03)	(0.04)	(0.06)	(0.06)
Controls	No	No	Yes	Yes
N	373	373	365	365
R2	0.1156	0.1557	0.1743	0.2179

Notes This table reports the results of a regression of the two measures of trust collected on the breadth of the (formal) rulebook. The first measure of trust, general trust, captures how much managers trust people *in general*. The second measure of trust, employee trust, captures how much managers trust *their own employees*. Finally, the interaction of these two terms is included. Standard errors are clustered at the firm-level, including two observations per firm. Standard errors are indicated in braces and significance is indicated as: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

I also ask respondents about their beliefs regarding good management through a series of hypothetical scenarios. For example, when hiring a candidate, should the decision be based on discretion or rules? Based on their responses, I create four indices: The *rules index* captures how respondents value rules over discretion. The *procedure index* measures the extent to which managers feel employees should follow procedures rather than show initiative. The *centralisation index* captures whether managers believe the CEO should be involved in decisions. The *consistency index* assesses whether managers believe that current decisions set a precedent for future decisions. Appendix 2.9.B details

these questions and the construction of these indices. Each of these metrics is generated by combining a set of Likert-scale responses through inverse covariance weighting.

I regress these indices on both the breadth of the formal rulebook. I find that managers who value consistency, rules, centralisation, and procedure more tend to use broader rules, showing the metric clearly correlates with managers' beliefs of good management. Table 2.6 shows that managers who value rules, procedure, centralisation and consistency work in firms with broad formal rulebooks. This suggests that either these managers have sufficient influence to affect the formal rules of their firm, or that their beliefs regarding good management are, in part, shaped by operating in a rule-based environment.

Table 2.6: The relationship between formal rules and managers' beliefs

	Formal Rule Breadth			
	(1)	(2)	(3)	(4)
Rules Index	0.136***			
	(0.03)			
Procedure Index		0.099***		
		(0.03)		
Delegation Index			0.158***	
			(0.03)	
Consistency Index				0.123***
				(0.03)
Constant	0.398***	0.398***	0.396***	0.396***
	(0.02)	(0.02)	(0.02)	(0.02)
Controls	No	No	No	No
N	384	384	384	384
R2	0.0791	0.0370	0.1153	0.0644

Notes This table displays the relationship between managers' beliefs and the formal rule breadth. The formal breadth variable captures the breadth of the formal rules only. The construction of the four indices is described in Appendix 2.9.B. The first measure relates to the importance of rules-based management practices, the second to the importance of following procedure, the third to the willingness to delegate to people, and the fourth to the importance of consistency. Standard errors, clustered at the firm level, are indicated in braces and significance is indicated as: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.4.3 Rules and management practices

In this section, I study how my measure of rule-based management relates to other dimensions of firm management. To study this, I first explore the relationship between management practices measured using MOPS-type questions (Bloom, Sadun, and Van Reenen, 2016b) and organizational rules. Next, I examine the relationship between rules and a set of HR management practices.

To study the relationship between the adoption of rules and structured management practices, I use a variation of the Management and Organisational Practices Survey (MOPS). Using these questions, I construct a measure of management quality based on these questions, specifically the management scores are calculated as the weighted average of z-scores, with weights determined by the inverse of the covariance matrix. I then regress this management score on the breadth of the full and formal rulebook. I present these results in Column (1) and (2) of Table 2.7. I find that the breadth of the rulebook including all rules does not correlate with the management score, but that the breadth of the rulebook including only formal rules strongly, and positively, correlated to the overall management scores.

One question in this survey focuses specifically on HR practices, asking: "When making a promotion decision, your decision is based mainly on...". For this question, a "good" answer focuses on the performance of the employee, whereas a "bad" answer focuses on rule-based promotion decisions based on experience and tenure. Unsurprisingly, firms with a broader formal rulebook tend to also have relatively rule-based system of making promotion decisions. More strikingly, in this context the response to this question is highly negatively correlated to the remaining questions. This suggests that rule-based promotions are seen as a best practice by these firms. I conjecture that in an environment where performance is difficult to accurately measure, such rule-based decisions can improve fairness and motivate employees to acquire position-specific human capital.

In columns three to six of Table 2.7, I focus separately on a management score excluding the HR management question, and a management score including *only* the

HR management question. Whereas both these scores are uncorrelated to the breadth of the full rulebook (including informal rules), they are respectively strongly positively, and strongly negatively, correlated to the breadth of the formal rulebook. This suggests that adopting rule-based HR management practices is indeed correlated with more documented rule-based management.[4]

Table 2.7: Rule breadth and MOPS management practices

	(1)	(2)	(3)	(4)	(5)	(6)
	Breadth	Formal breadth	Breadth	Formal breadth	Breadth	Formal breadth
	b/se	b/se	b/se	b/se	b/se	b/se
Management Score	0.010	0.061**				
	(0.02)	(0.03)				
Management Score			0.013	0.084***		
Excluding HR			(0.02)	(0.03)		
Management Score					0.011	-0.068***
Only HR					(0.02)	(0.02)
Constant	0.614***	0.399***	0.614***	0.399***	0.614***	0.402***
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)
N	384	384	384	384	368	368
R2	0.0009	0.0232	0.0014	0.0449	0.0010	0.0284
Adj-R2	-0.0018	0.0206	-0.0013	0.0424	-0.0017	0.0258

Notes This table displays the results from regressions of the (formal) rule breadth on four dimensions of manager beliefs. The MOPS score is the inverse covariance weighted average z-score of all responses to the management questions. MOPS excluding HR and MOPS HR only implements the same procedure including respectively all questions except the HR question, and only the HR question. Standard errors are clustered at the firm-level, including two observations per firm. Standard errors are indicated in braces and significance is indicated as: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A more detailed examination of the link between HR rules and HR practices reveals an interesting picture. Broadly speaking, managers with broader rulebooks appear to be more active managers. Table 2.8 shows that HR managers with broader rules identify more problems related to turnover, lack of skills in potential employees, and lack of motivation. Moreover, although not shown here, as a proportion of their total number of

[4] Note that the management scores excluding the HR practice question and including only the HR practice question are strongly negatively correlated.

employees, these managers have hired and expect to hire more people, and have terminated a larger share of their workforce in the last 12 months. They spend significantly more when hiring a professional employee and also invest more when hiring other employees. They also give pay rises more frequently. When advertising job openings, firms with more rules are more likely to place advertisements in a gazette or on Ethiojobs or other vacancy websites but are less likely to use a recruitment agency or post advertisements on university campuses. These relationships remain robust after controlling for other firm observables. Overall, it appears that firms with more formal rules engage in more active HR management than those with fewer rules.

Table 2.8: The relationship between the rule breadth and human resource problems as identified by managers.

	Lack of motivation	High turnover	Lack of skills	Number of problems
Rule Breadth	0.234**	0.257***	0.378***	0.869***
	(0.09)	(0.09)	(0.09)	(0.22)
Constant	0.312***	0.303***	0.244***	0.859***
	(0.07)	(0.07)	(0.06)	(0.16)
N	386	386	386	386
R2	0.0267	0.0324	0.0695	0.0606

Notes This table displays regression results linking the breadth of (formal) organisational rules to HR problems identified by managers. The first three dependent variables are dummy indicators for three HR issues: lack of motivation, high turnover, and skills shortage. The fourth dependent variables is the total number of problems identified. Standard errors are indicated in braces and significance is indicated as: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.5 Rule-based management and firm performance

To validate the measure of rules developed in this paper, it is important to consider how these rules correlate with firm performance, and whether they help explain differences in performance across otherwise similar enterprises. This supports the external validity of the measure as being related to actual firm outcomes.

2.5.1 Rules and profitability

To assess these relationships, I use profit data collected from a subset of 116 firms in an exercise similar to that conducted in Bloom and Van Reenen (2007). I regress a number of measures of firm profitability on the breadth of the formal rulebook. Specifically, I estimate the following regression:

$$y_{fs} = \alpha + x_{1fs} \cdot \beta_1 + \mathbf{X}'_{2f}\boldsymbol{\beta}_2 + \epsilon_{fs} \tag{2.3}$$

where y_{fs} is the specific measure of performance used for firm f in scenario s. x_{1fs} is the breadth of the rulebook, and β_1 thus captures the conditional correlation between rule-based management and the outcome of interest. The control vector $\boldsymbol{X}_{2f}$ includes the number of employees, the MOPS management score, whether the plant is a subsidiary, the sector in which the firm operates, and the type of employee discussed. As I have two observations of rule breadth per firm I cluster standard errors at the firm level.

I run this regression for three measures of firm performance: sales, profit, and profit over sales (the profit ratio). Table 2.9 reports the results from this regression including the full set of controls. Column (1) shows that firms with more rules in general do not have more sales conditional on covariates, however column (2) shows that these firms do make on average more profit, and as shown in column (3) these firms have a higher ratio of profit over sales. These effects are economically meaningful, with a firm with a rule breadth of zero earning around 60% of the profit and profit ratio of a firm with a rule breadth of one.

Next, I turn to the relationship between formal rules and firm performance. Column (4) shows that the use of formal rules is positively associated with the sales, and that the adoption of formal rules is more positively correlated to profitability than the full set of rules (column 5). The link between formal rules and the profit over sales ratio is weaker than between the full rulebook and profit over sales, likely due to the positive link between this measure of rule breadth and sales. These results suggest that there is a large and economically meaningful link between the adoption of rules by managers and

firm performance.

Table 2.9: Organisational rules and firm performance

	(1) Sales	(2) Profit	(3) Profit/Sales	(4) Sales	(5) Profit	(6) Profit/Sales
Breadth	27.844	4668.002*	0.123***			
	(38.14)	(2382.13)	(0.05)			
Breadth formal				70.720*	5373.005**	0.081*
				(39.12)	(2413.71)	(0.04)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. var.	90.7	6515.7	0.2	90.7	6515.7	0.2
N	242	232	230	242	232	230
R2	0.200	0.119	0.100	0.215	0.130	0.082
Adj-R2	0.172	0.087	0.067	0.188	0.099	0.049

Notes This table displays the relationship between firm performance and rule breadth. The rule breadth variable captures the span of formal rules and informal rules taken together, whereas the formal rule breadth only includes the formal rules. The three dependent variables included are sales in the past month in millions of Birr, profit in the past month in thousands of Birr, and the ratio of profit over sales. These variables are all winsorised at the 95th percentile. The included controls are the number of employees, sector, the type of position, management practices, and whether the firm is a subsidiary. All standard errors are clustered at the firm level. Figure 2.1 details these variables, the number of employees is incorporated as a continuous variable rather than by quartile. Standard errors are indicated in braces and significance is indicated as: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

One concern is that the firms that choose to adopt formal rules are simply different on other fronts, for example in terms of size, industry or management practices. To address this concern, focusing on the link between formal rules and performance, I progressively add individual components of the vector X_2 and assess how the results respond to this inclusion. I report the results from this exercise in Table 2.10.

Table 2.10: Explaining firm performance with organisational rules

Panel A: Sales						
	(1)	(2)	(3)	(4)	(5)	(6)
Breadth formal	103.450**	74.754*	81.720**	69.762*	70.677*	70.661*
	(40.26)	(38.19)	(37.88)	(37.82)	(38.33)	(38.97)
Unc. mean dep. var.	90.716	90.716	90.716	90.716	90.716	90.716
N	246	246	246	246	242	242
Adj-R2	0.039	0.190	0.199	0.202	0.195	0.188
Panel B: Profit						
	(1)	(2)	(3)	(4)	(5)	(6)
Breadth formal	7.601***	6.627**	6.832**	6.864**	6.682**	6.632**
	(2.65)	(2.91)	(2.98)	(2.82)	(2.85)	(2.85)
Unc. mean dep. var.	6.028	6.028	6.028	6.028	6.028	6.028
N	236	236	236	236	232	232
Adj-R2	0.049	0.069	0.070	0.066	0.069	0.063
Panel C: Profit over sales						
	(1)	(2)	(3)	(4)	(5)	(6)
Breadth formal	0.051	0.071	0.074	0.078*	0.082*	0.083*
	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.04)
Unc. mean dep. var.	0.195	0.195	0.195	0.195	0.195	0.195
N	234	234	234	234	230	230
Adj-R2	0.004	0.038	0.036	0.032	0.037	0.055
Firm Size	No	Yes	Yes	Yes	Yes	Yes
Management Practices	No	No	Yes	Yes	Yes	Yes
Subsidiary	No	No	No	Yes	Yes	Yes
Position	No	No	No	No	Yes	Yes
Industry	No	No	No	No	No	Yes

Notes This table examines the robustness of the relationship between firm sales, profitability, and profit over sales with formal rule breadth. All three dependent variables are winsorised at the 95th percentile. The included controls are the number of employees, sector, the type of position, management practices, and whether the firm is a subsidiary. Figure 2.1 details these variables, the number of employees is incorporated as a continuous variable rather than by quartile. Standard errors, clustered at the firm-level, are indicated in braces and significance is indicated as: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Panels A and B of Table 2.10 suggest that the relationship between formal rule

adoption and both sales and profit remains robust when covariates are included. For both outcomes the relationship is both the largest, and most statistically significant without additional controls. For both of these covariates however the results remain statistically significant at respectively the 10% and 5% level throughout columns one to six. The relationship between the adoption of rules and profit over sales is positive throughout the specifications, but only robust once I control for firm size, the management score and whether the plant is a subsidiary of a larger firm (column 4, panel C).

2.5.2 Explaining performance differences between firms

Part of the motivation for this empirical work are the persistent performance differences between seemingly similar enterprises. An important question is therefore to what extent these measures of rules can help explain these differences. To test this, I compare the performance of the MOPS score with the performance of my measure of rules in explaining differences in performance across firms, conditional on the vector of controls X_2.

As noted previously, the patterns in the adoption of human resource management practices relative to other management practices are striking. Firms that are relatively well-managed in terms of their other practices tend to have particularly rule-based promotion practices, which are generally regarded as "bad" practices. To aid the comparison, I drop this measure of promotion practices from my management index. I report the results including this in Appendix Table 2.18, noting including this variable this strengthens the argument made in this section.

Columns (1) and (2) of Table 2.11 show that, relative to regressing the three measures of performance on a constant, the effect of including the MOPS score on the adjusted R-squared is ambiguous, with the adjusted R-squared falling in two of the three specifications. Focussing on columns (1) and (3), I find that instead including the measure of the formal rule breadth clearly increases the adjusted R-squared in all three specifications by between 11% (Panel A) and 45% (Panel B). Finally, focussing on columns (3) and (4), also including the management score reduces the adjusted R-squared in all three specifications.

Table 2.11: Management practices, rules and firm performance

	Panel A: Sales			
	(1) Sales	(2) Sales	(3) Sales	(4) Sales
Management Score		-15.104		-18.993
(Excluding HR)		(20.15)		(20.70)
Breadth formal			64.962	72.329*
			(39.32)	(38.94)
N	242	242	242	242
Adj-R2	0.1669	0.1691	0.1848	0.1838
	Panel B: Profit			
	Profit	Profit	Profit	Profit
Management Score		397.719		135.237
(Excluding HR)		(504.41)		(561.65)
Breadth formal			5257.050**	5225.160**
			(2340.02)	(2410.80)
N	232	232	232	232
Adj-R2	0.0752	0.0721	0.1049	0.0965
	Panel C: Profit over sales			
	Profit/Sales	Profit/Sales	Profit/Sales	Profit/Sales
Management Score		0.014		0.010
(Excluding HR)		(0.02)		(0.02)
Breadth formal			0.080*	0.076*
			(0.04)	(0.04)
N	230	230	230	230
Adj-R2	0.0364	0.0361	0.0518	0.0497

Notes This table examines how predictive the formal rule breadth is for sales, profit and profit over sales relative to the MOPS score excluding the HR question. The calculation of the management score and formal breadth are detailed earlier in this section. All standard errors are clustered at the firm level. The included controls are the number of employees, sector, the type of position, management practices, and whether the firm is a subsidiary. Figure 2.1 details these variables, the number of employees is incorporated as a continuous variable rather than by quartile. Standard errors, clustered at the firm-level, are indicated in braces and significance is indicated as: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.6 Rules and resilience

Existing research establishes a connection between a firm's resilience to economic shocks and the autonomy granted to its managers and employees (Aghion et al., 2021; Li et al., 2022). In particular, Li et al. (2022) develop a theoretical model in which economic shocks cause relational contracts to become more rule-based, and these new rules then reduce both long-term performance and resilience to future shocks. As the survey data was collected at the end of 2021, I included questions on the effect of COVID-19 on firms to test the hypothesis that rule-based management reduces resilience to shocks.

I find that firms with a broader rulebook were significantly more affected by the economic shock, and that this result is robust to the inclusion of additional controls. I regress four metrics on the breadth of the full and formal rulebook: (1) whether cash flow decreased due to COVID-19, (2) whether the firm had to postpone purchases due to the pandemic, (3) whether the firm experienced financial problems, and (4) whether profit decreased in the financial year following the pandemic compared to the financial year ending shortly after the first lockdown. Throughout, I include the same set of controls as in the regressions related to firm profitability.

In Panel A of Table 2.12, I focus on the link between the full set of organizational rules adopted by the manager and the four outcome variables. I find that adopting a more rule-based management style is positively correlated with the probability of experiencing all four negative effects. This suggests that although firms with more rules remain more profitable as shown in Table 2.9, they were also more negatively affected by the COVID-19 restrictions imposed in Ethiopia.

To provide initial evidence on whether the adoption of these rules is a response to the exposure to the economic shock, or whether firms with more rules were more strongly affected, I implement two further tests. First, the results from Panel B suggest that firms with more formal rules were not consistently more affected by COVID-19, supporting the idea that firms adopted informal rules in response to the shock, rather than the shock having heterogeneous effects depending on pre-existing rule adoption in line with Li et al.

(2022).

Table 2.12: Relationship breadth rulebook and firm resilience

	Panel A: All rules and resilience			
	(1) Decrease cash flow	(2) Postpone purchases	(3) Financial problems	(4) Fall profit
Breadth	0.264***	0.258***	0.417***	0.314***
	(0.08)	(0.07)	(0.09)	(0.11)
Constant	0.719***	0.080	0.555***	0.194*
	(0.08)	(0.08)	(0.10)	(0.11)
Controls	Yes	Yes	Yes	Yes
N	376	376	376	230
R2	0.1114	0.0881	0.1374	0.1006
	Panel B: Formal rules and resilience			
	Decrease cash flow	Postpone purchases	Financial problems	Fall profit
Formal Breadth	0.052	0.161**	0.113	0.034
	(0.07)	(0.07)	(0.08)	(0.11)
Constant	0.868***	0.195***	0.782***	0.374***
	(0.06)	(0.07)	(0.08)	(0.11)
Controls	Yes	Yes	Yes	Yes
N	376	376	376	230
R2	0.0716	0.0679	0.0641	0.0550

Notes: This table examines the relationship between the breadth of the identified rules and the effects of Covid-19 restrictions on the firm. Decrease cash flow, postpone purchases and financial problems equal one if the manager indicates they had any of these issues due to Covid-19 restrictions. The indicator fall profit equals one if the firm reports lower profit in the 2020-2021 financial year than the 2019-2020 financial year. The included controls are the number of employees, sector, the type of position, management practices, and whether the firm is a subsidiary. Figure 2.1 details these variables, the number of employees is incorporated as a continuous variable rather than by quartile. Standard errors, clustered at the firm level, are indicated in braces and significance is indicated as: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Finally, in a more exploratory test, I expand the dataset using survey data collected from a subset of the firms in 2017. As I do not have data on rules for this period, I use the

adoption of rule-based *management practices* as a proxy. Table 2.13 gives the results from this exercise, which I run without controls due to data availability. Contemporaneous rule-based management (MOPS) is positively correlated with both higher crisis-era profits and a larger post-crisis profit decline, whereas past rule-based practices show no such relationship.

Table 2.13: Relationship breadth rulebook and firm resilience

	Panel A: Current rules-based HR practices and resilience			
	(1) Decrease cash flow	(2) Postpone purchases	(3) Financial problems	(4) Fall profit
MOPS HR Score	-0.024	0.023	0.067**	0.094**
	(0.03)	(0.03)	(0.03)	(0.04)
Constant	0.780***	0.215***	0.635***	0.393***
	(0.03)	(0.03)	(0.03)	(0.04)
N	200	200	200	123
R2	0.0035	0.0032	0.0193	0.0406
	Panel B: Past rules-based HR practices and resilience			
	Decrease cash flow	Postpone purchases	Financial problems	Fall profit
MOPS HR Score	-0.034	0.020	0.013	-0.058
	(0.03)	(0.04)	(0.05)	(0.06)
Constant	0.770***	0.195***	0.588***	0.307***
	(0.04)	(0.04)	(0.05)	(0.06)
N	113	113	114	68
R2	0.0064	0.0027	0.0006	0.0174

Notes This table examines the relationship between the rules-based HR practices and the effects of Covid-19 restrictions on the firm. Decrease cash flow, postpone purchases and financial problems equal one if the manager indicates they had any of these issues due to Covid-19 restrictions. The indicator fall profit equals one if the firm reports lower profit in the 2020-2021 financial year than the 2019-2020 financial year. The included controls are the number of employees, sector, the type of position, management practices, and whether the firm is a subsidiary. Figure 2.1 details these variables, the number of employees is incorporated as a continuous variable rather than by quartile. Robust standard errors are indicated in braces and significance is indicated as: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.7 Assessing the performance of the sampling algorithm

This section assesses the performance of the Bayesian adaptive algorithm through three analyses: (1) analysing data collected from firm managers in the pilot, (2) using simulated data to identify settings where the algorithm performs well, and (3) comparing the methodology to that of Chapman et al. (2024) in measuring the parameters of a utility function.

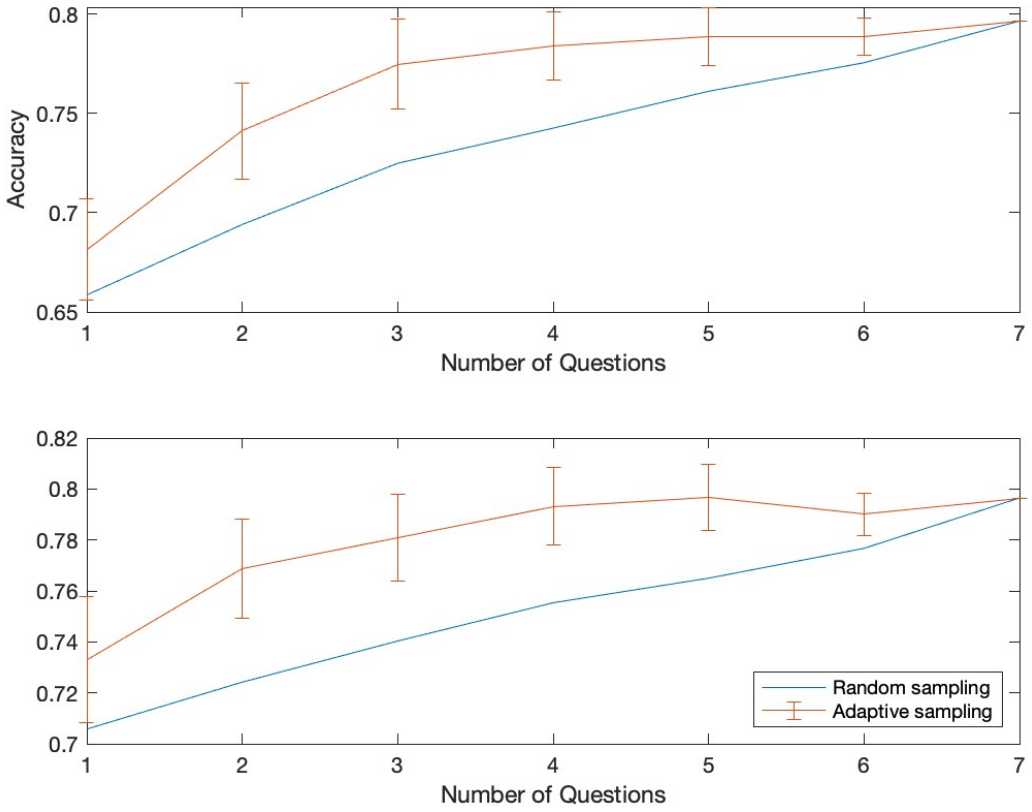
2.7.1 Measuring managerial decisions using adaptive sampling

I assess whether the adaptive questionnaire outperforms random question selection by using data from a different sample of 187 managers. Each manager made decisions for eight randomly selected cases from the scenario landscape. I use this data to test the adaptive algorithm against random sampling. Appendix 2.9.D reports the estimated parameters for the GP based on this sample used in the analysis and data collection.

To compare the performance of the adaptive algorithm and random sampling, I use training data to conduct an exercise where each respondent's eight observations are split into seven training cases and one test case repeatedly. I then compare the performance of adaptive versus random sampling in predicting the value of that test case after each sampling step. The adaptive sampling algorithm's accuracy – the average probability the algorithm makes a correct prediction – is 6 percentage points higher for the pay rise scenario and 5 percentage points higher for the hiring scenario after three questions. Figure 2.3 depicts the results from a clustered OLS regression comparing random and adaptive sampling.

Table 2.14 depicts the performance of adaptive relative to random sampling. The top two rows of results depict the efficiency ratio of the adaptive sampling algorithm. Efficiency ratio is defined as N/N^*—the number of questions a random design (N) needs to reach a given accuracy divided by the number the adaptive design (N^*) needs; values greater than 1 indicate a gain from using the adaptive algorithm. The sampling algorithm increases the learning rate by 60% to 80% over 7 questions.

Figure 2.3: Assessing the performance of the adaptive algorithm using training data



Notes This figure compares the performance of randomly sampling observations (the blue line) with sampling them adaptively (in red). The set of responses is repeatedly split in a training set (7 cases) and test set (1 case) for each respondent to implement this. The confidence bands for the performance of the adaptive sampling algorithm are based on standard errors clustered at the (simulated) individual level. The accuracy is the average probability that the algorithm makes a correct prediction for the test case.

However, in this exercise the adaptive sampling algorithm is heavily constrained, as it can only choose from the seven randomly sampled questions rather than the full set of 128 cases. To evaluate the algorithm's potential without this constraint, I follow the literature on optimal experimental design (Chapman et al., 2024). Specifically, I train the model using the full dataset, to then simulate individuals for whom we have observed the full set of cases using the trained prior and the observed training data to generate $f(X_o, Y_o)$. The final two rows of results in Table 2.14 show the adaptive algorithm now increases the rate of learning by 50 to 300% compared to randomly sampling questions (depending on the number of questions asked). As described in section 2.7.2, the algorithm does particularly well when there is a lot of structure to exploit, which is the case in the pay rise scenario.

Table 2.14: Efficiency Ratio training sample

	Constrained Efficiency Ratio		
Scenario	2 Questions	3 Questions	4 Questions
Pay rise scenario	1.57	1.66	1.73
Hiring scenario	1.78	1.76	1.77
N	198	198	198
	Unconstrained Efficiency Ratio		
Scenario	5 Questions	10 Questions	15 Questions
Pay rise scenario	1.57	2.57	4.12
Hiring scenario	1.56	2.17	2.59

Notes Table 2.14 depicts the efficiency ratio of the adaptive sampling algorithm relative to random sampling using the training data. For each respondent, the answers are repeatedly randomly split in a training set (7 cases) and test set (1 case). The adaptive and random sampling algorithm are applied to the training set. I generate a benchmark based on the average accuracy of the two algorithms after 2, 3 and 4 questions. I then determine the first point where each of these algorithms achieves this accuracy to calculate, respectively N and N^* for random and adaptive sampling. I calculate the efficiency gains as the mean of N over the mean of N^* to create a symmetric measure. The constrained efficiency ratio uses only the real data. The unconstrained efficiency ratio uses the observed data to simulate the full set of observations. To do so, I train the hyperparameters of the Gaussian Process, specifically a linear mean function and automatic relevance detection square exponential kernel. I use this to generate data based on the observed data for each individual and this prior trained on the full training sample. In predicting managers' decisions using real data, the adaptive algorithms achieves 60-80% faster learning than random sampling. In the simulations using the data that was simulated using the real data the adaptive sampling algorithm massively outperforms random sampling.

This section shows the algorithm significantly increases our rate of learning rela-

tively to randomly sampling questions. Based on these results, I would expect that the accuracy after eight adaptively sampled questions is approximately the same as after 15 to 20 randomly sampled questions.

2.7.2 Analysis using simulated data

In this section, I study the performance of the Bayesian adaptive questionnaire using simulated data. These simulations help identify situations where the adaptive algorithm performs particularly well. I simulate data from various Gaussian Processes (GPs) and compare the performance of the adaptive sampling algorithm to random sampling. The simulations show that the adaptive algorithm significantly increases the rate of learning, especially when it can exploit the structure in the relationship between outcomes Y and characteristics X. This is evident when different characteristics contribute differently to the covariance and when the mean function provides information about predictability. Additionally, the more questions we ask a respondent, the better the algorithm becomes at selecting informative questions based on prior responses.

To simulate the data, I use a squared exponential kernel with automatic relevance detection (ARD). In this specification, each case $\mathbf{x}_i$ has D binary characteristics, and l_d is the length-scale parameter for characteristic d. In this kernel, the covariance between two cases decreases exponentially with the squared distance between their characteristics, where the length-scale parameter l_d controls sensitivity to differences in each characteristic.

$$k(\mathbf{x}_p, \mathbf{x}_q) = \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_{pd} - x_{qd})^2}{l_d^2}\right) \tag{2.4}$$

Each data-generating process (DGP) has six binary characteristics. Generally, l_d is set to 1, except where explicitly stated otherwise. For process 1, l_d is set to 1 for all characteristics. In process 2, I set l_1 to a very small value, which means that the covariance between cases that differ in characteristic 1 drops off to nearly zero, effectively treating those cases as uncorrelated in this dimension. For processes 3 and 4, I set one

Table 2.15: The eight data-generating processes used for simulations

	No Mean Function	Mean Function
l_1 equals 1	DGP 1	DGP 5
l_1 is very low	DGP 2	DGP 6
l_1 is high	DGP 3	DGP 7
l_1 and l_2 are high	DGP 4	DGP 8

Notes: The data-generating processes all have six binary characteristics that determine a binary outcome. The covariance is modeled using an automatic relevance detection squared exponential kernel. Unless otherwise noted, the length-scale parameters for the remaining characteristics are set to 1. The mean function is a simple linear mean with some random noise added.

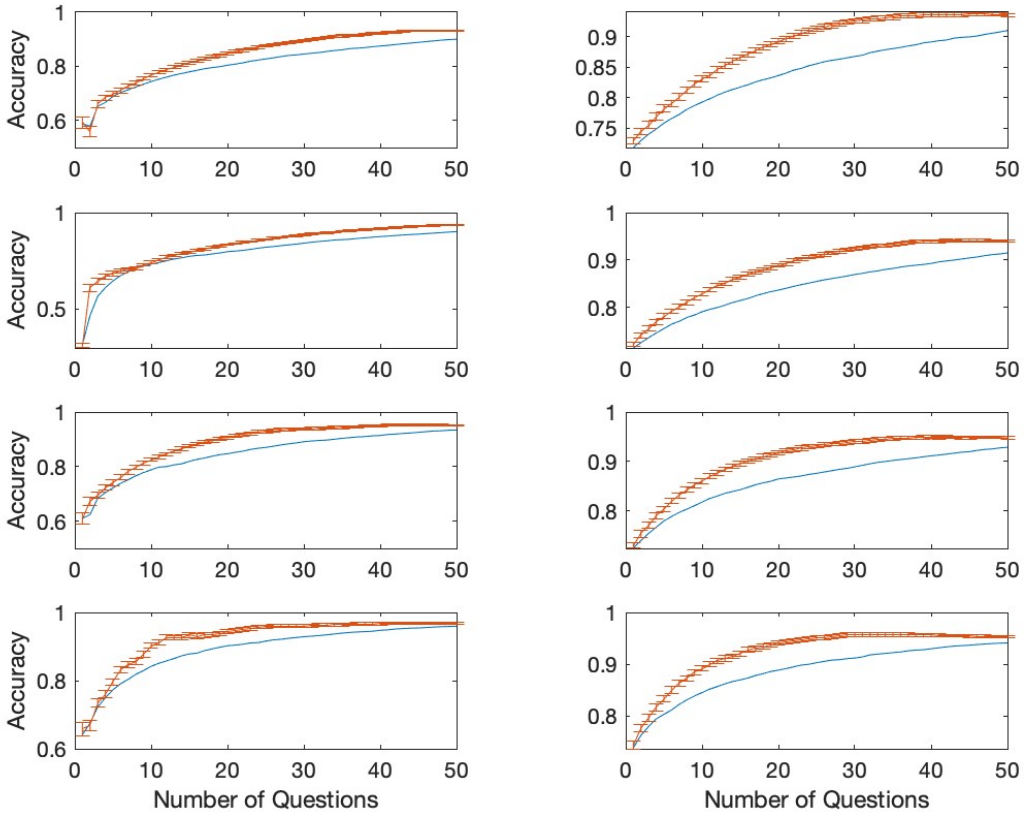
and two length-scale parameters respectively to high values, meaning that differences in those specific characteristics have little effect on the covariance between cases. Processes 5-8 use the same four combinations of length-scale parameters, but with a linear mean function added to the Gaussian Process. Table 2.15 summarizes these processes.

In Figure 2.4, I compare the adaptive Bayesian questionnaire to randomly sampling a sequence of questions. This figure shows that as the structure on the problem increases, the adaptive sampling algorithm can leverage this information to further outperform random sampling. All these problems have sufficient structure for adaptive sampling to significantly outperform random sampling. Detailed regression results relating to the performance gains can be found in Table 2.17 in Appendix 2.9.A.

While the absolute gains in accuracy may appear modest, a more meaningful metric is the efficiency ratio, calculated as the ratio of the number of questions required by random sampling (N) to achieve the same level of information as the optimized adaptive sampling (N^*), following (Chapman et al., 2024). Table 2.16 presents these efficiency ratios:

The results indicate that adaptive sampling requires significantly fewer questions to achieve the same level of accuracy as random sampling. The efficiency gains range from 15% to 120%, depending on the DGP. The gains are larger when the problem has more structure that the adaptive algorithm can exploit. One might argue that random sampling is a naive benchmark. However, alternative strategies, such as sampling cases that are very different from each other, may not perform better. In fact, such strategies can worsen performance compared to random sampling, as shown in Appendix Figure

Figure 2.4: Assessing the performance of the adaptive algorithm using simulated data



Notes: This figure compares the performance of randomly sampling observations (the blue line) with sampling them adaptively (in red). The confidence bands for the performance of the adaptive sampling algorithm are based on standard errors clustered at the (simulated) individual level. The accuracy is the average probability that the algorithm makes a correct prediction for each of the decision problems. The specific DGPs used are described in Table 2.15.

Table 2.16: Efficiency ratio of adaptive versus random sampling

DGP	Efficiency Ratio after 5 Questions	10 Questions	15 Questions
DGP 1	1.16	1.26	1.33
DGP 2	1.37	1.23	1.28
DGP 3	1.48	1.38	1.48
DGP 4	1.30	1.59	1.59
DGP 5	1.78	1.80	2.01
DGP 6	1.76	1.84	1.89
DGP 7	1.87	2.05	2.10
DGP 8	1.80	2.13	2.19

Notes This table reports the efficiency ratio of the adaptive sampling algorithm relative to random sampling. For each randomly generated individual, the adaptive and random sampling algorithm are applied. I generate a benchmark based on the average accuracy of the two algorithms after 5, 10 and 15 questions. I then determine the first point where each of these algorithms achieves this accuracy to calculate N and N^*. I calculate the efficiency gains as the mean of N over the mean of N^* to create a symmetric measure. Here, the adaptive algorithms achieves 15-120% faster learning than random sampling depending on the structure of the problem.

2.5.

These simulations demonstrate that the Bayesian adaptive questionnaire significantly outperforms random sampling in identifying informative questions, especially when the underlying decision process has exploitable structure. This efficiency is valuable in practical applications where the number of questions must be limited to reduce respondent burden.

2.7.3 Measuring Preferences

To further assess the performance and flexibility of the Bayesian adaptive questionnaire, I compare it to the Dynamic Optimisation of Sequential Experiments (DOSE) algorithm developed by Chapman et al. (2024). The DOSE algorithm selects questions by maximising the expected information gain about the parameters, measured by the Kullback-Leibler divergence. Their methodology is applied to differentiate between models and estimate the parameters of a prospect-theory utility function. In their application, respondents are asked to choose between a lottery and a certain payoff. The utility model

includes three parameters: ρ, the coefficient of relative risk aversion, λ capturing loss aversion, and μ which captures the attentiveness of the respondent. The specific utility function is:

$$u(x) = \begin{cases} w^{\rho}, & \text{if } w \geq 0 \\ -\lambda(-w)^{\rho}, & \text{if } w < 0 \end{cases} \tag{2.5}$$

where, denoting by x and z the payoffs of the lottery and y the payoff of an alternative certain payoff.

To implement the Bayesian adaptive questionnaire in this context, I incorporate the same model structure. After each question, I update the posterior distribution of the parameters based on the respondent's answer and use this updated distribution to inform the selection of the next question. The mean function and kernel are updated accordingly.

The set of possible questions involves lotteries where the certain payoff y is set to zero, and the lottery payoffs x (positive) range from 0.25 to 5 in increments of 0.25, while z (negative) ranges from -5 to 0.5 in increments of 0.5, resulting in 200 possible questions. This setup mirrors that of Chapman et al. (2024).

I compare the performance of my algorithm to theirs in estimating the three parameters. To do so, I simulate data for 200 individuals, I find no significant difference in the performance of the two algorithms. After thirty questions, the Bayesian Adaptive Questionnaire marginally outperforms DOSE in estimating two of the three parameters; after one-hundred questions the roles are reversed with DOSE performing marginally better estimating two of the three parameters. None of these results have any statistical significance.

This comparison shows that the Bayesian adaptive questionnaire matches the performance of specialized algorithms like DOSE in this setting. The key advantage of the Bayesian adaptive questionnaire is its flexibility; it can be applied to a wide range of problems without substantial modifications. This makes it a valuable tool for various applications in experimental design and survey methodology, beyond the specific context of measuring preferences. This comes at the cost of an increased computational

requirements.

2.8 Discussion

Explaining persistent performance differences between seemingly similar enterprises is fundamentally important for designing and targeting policy. The question of what managers do, and what makes managers effective, is of first-order importance to firms in low-income countries. Rules were fundamental to unlocking economies of scale during the development of modern U.S. manufacturing (Taylor, 1911); yet, we know very little about how managers use rules today. We know that firms in sub-Saharan African countries reap limited benefits from economies of scale (Bassi et al., 2023), and that firms in low-income countries face high costs of delegation (Akcigit et al., 2021a; Bloom et al., 2013a). The methodology developed here allows us to measure a dimension of organisational rules that appears to matter for firm performance, enabling the study of a wide range of questions about managerial practices and their impact on firm outcomes.

I argue that we know so little about rules because measuring such a high-dimensional construct is difficult using traditional economic methodologies. To address this challenge, I have developed a novel Bayesian adaptive questionnaire to measure organisational rules. This tool efficiently manages the high dimensionality by sampling the most informative questions. I implemented this methodology to survey a set of large Ethiopian firms. The results show that having a broader formal rulebook is positively correlated with firm performance and can help explain heterogeneity in performance between firms. These findings are robust to the inclusion of additional firm observables as explanatory variables. However, in line with theoretical results from Li et al. (2022), this increased performance appears to come at the cost of greater vulnerability to economic shocks. Firms that have more rules perform better but report being more negatively affected by the COVID-19 restrictions implemented in Ethiopia in 2020.

This suggests that implementing formal rules might enhance the performance of relational contracts within firms but comes at the cost of increased vulnerability to changes in the economic environment. This aligns with the findings of Aghion et al. (2021), which

show that reduced autonomy—and thus flexibility—at the plant level diminishes firms' resilience to economic turbulence. Crucially, the results in this paper demonstrate that organisational rules are a significant factor in explaining heterogeneity in performance across firms.

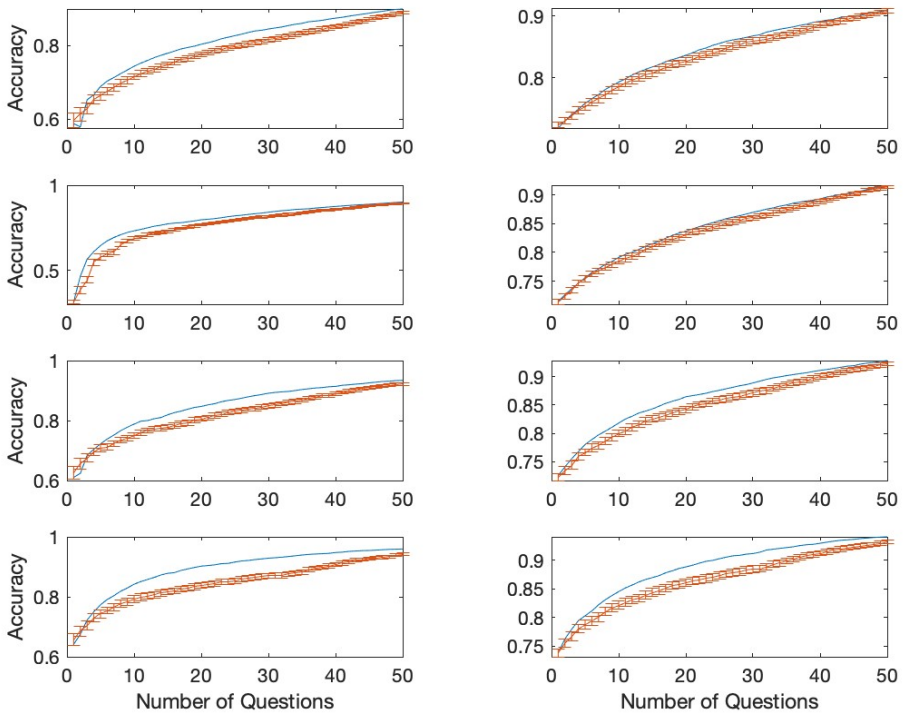
Understanding how managers use rules is fundamentally important, particularly in low-income countries where managerial span of control appears to be a key constraint (Bassi et al., 2023; Bloom, Mahajan, McKenzie, and Roberts, 2010). The development of a feasible tool to measure rules opens up a number of avenues for policy-relevant research. First, using experimental exogenous variation to promote the implementation of formal rules could assess the causal link between rules and firm performance, with a strong focus on heterogeneity in treatment effects. Second, studying the shared understanding and enforcement of rules within a firm would provide insights into internal dynamics. Third, intersecting with the first two, examining how rules affect a firm's ability to react to changes—and whether the lower resilience to COVID-19 observed in this paper reflects a reduction in flexibility—would be valuable. Throughout this research, focusing on the theoretical modelling of rules in relational contracts is important to understand how rules affect outcomes.

The second contribution of this paper is the development of a novel method of questionnaire design. This method is broadly applicable and outperforms alternative generic sampling algorithms. Compared to random sampling, this algorithm doubles the rate of learning. The application presented in this paper demonstrates that the method can be feasibly implemented in a field setting. This approach has broad future applications in survey design where asking questions is costly. For example, applying this method to phone surveys is a potential next step. This Bayesian adaptive questionnaire methodology has the potential to significantly reduce the cost, both for interviewers and respondents, of collecting data for researchers and policymakers.

2.9 Appendices

2.9.A Performance of the adaptive sampling algorithm

Figure 2.5: Assessing the performance of adaptive versus dispersed sampling



Notes: This figure compares the performance of sampling a dispersed set of observations (the blue line) with sampling them adaptively (in red). The confidence bands for the performance of the adaptive sampling algorithm are based on standard errors clustered at the (simulated) individual level. The accuracy is the average probability that the algorithm makes a correct prediction for each of the decision problems. The dispersed questions are sampled by determining at each step which unobserved case has the fewest shared characteristics with the set of observed cases. The specific DGPs used are described in Table 2.15.

Table 2.17: Gains in accuracy from adaptive sampling

DGP		(1) Step 5	(2) Step 10	(3) Step 15	(4) Step 25
1	Average increase due to adaptive sampling	0.00431	0.00928*	0.0345***	0.0478***
		(0.0107)	(0.00497)	(0.00271)	(0.00235)
	Average accuracy for random sampling	0.588***	0.689***	0.780***	0.827***
		(0.00885)	(0.00559)	(0.00385)	(0.00286)
2	Average increase due to adaptive sampling	-0.00656	0.0388***	0.0211***	0.0406***
		(0.00584)	(0.00601)	(0.00299)	(0.00229)
	Average accuracy for random sampling	0.315***	0.645***	0.773***	0.820***
		(0.00453)	(0.00619)	(0.00356)	(0.00275)
3	Average increase due to adaptive sampling	-0.00119	0.0186***	0.0550***	0.0589***
		(0.0111)	(0.00509)	(0.00326)	(0.00269)
	Average accuracy for random sampling	0.610***	0.722***	0.819***	0.870***
		(0.00966)	(0.00599)	(0.00389)	(0.00296)
4	Average increase due to adaptive sampling	0.0167	0.0225***	0.0554***	0.0423***
		(0.0104)	(0.00508)	(0.00347)	(0.00258)
	Average accuracy for random sampling	0.642***	0.773***	0.878***	0.916***
		(0.0103)	(0.00618)	(0.00373)	(0.00279)
5	Average increase due to adaptive sampling	0.0109***	0.0238***	0.0485***	0.0575***
		(0.00250)	(0.00289)	(0.00227)	(0.00204)
	Average accuracy for random sampling	0.719***	0.759***	0.818***	0.856***
		(0.00553)	(0.00482)	(0.00350)	(0.00271)
6	Average increase due to adaptive sampling	0.00797***	0.0265***	0.0490***	0.0533***
		(0.00200)	(0.00270)	(0.00215)	(0.00185)
	Average accuracy for random sampling	0.715***	0.756***	0.815***	0.855***
		(0.00522)	(0.00477)	(0.00375)	(0.00291)
7	Average increase due to adaptive sampling	0.00719	0.0352***	0.0502***	0.0514***
		(0.00636)	(0.00536)	(0.00538)	(0.00463)
	Average accuracy for random sampling	0.747***	0.794***	0.856***	0.890***
		(0.0127)	(0.0109)	(0.00863)	(0.00605)
8	Average increase due to adaptive sampling	0.00531*	0.0244***	0.0528***	0.0551***
		(0.00318)	(0.00277)	(0.00230)	(0.00207)
	Average accuracy for random sampling	0.725***	0.780***	0.843***	0.876***
		(0.00590)	(0.00493)	(0.00363)	(0.00288)

Table 2.17 gives regression results for the simulations in section 2.7.2, and specifically figure 2.4. The difference in accuracy between random and adaptive sampling is evaluated after 5, 10, 15 and 25 questions. The constant is the average predictive accuracy of random sampling, and the parameter on adaptive is the treatment effect of adaptive sampling. Standard errors are clustered at the simulated individual level and indicated in parentheses. .

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.9.B Construction of the indices

We ask respondents questions about their preferences over management practices in four hypothetical scenarios. The introduction of the four scenarios is as follows:

1. A manager is looking to hire a new employee for the main operation of the firm he works at. When recruiting new employees:

2. A production worker comes to a manager asking for a pay rise, when dealing with this request:

3. A customer has bought something from the firm on credit, and comes to the manager saying he cannot pay before the deadline. When dealing with this customer:

4. An employee comes to the manager with a complaint about another employee. When dealing with such a complaint:

For each of these scenarios, the respondents are asked, variations of the following questions on a Likert scale. These are always interpreted as a higher number representing less flexibility:

1. It is more important to hire the right person than to follow a clear recruitment procedure.

2. It is more important to find a new employee who shows initiative than someone who will follow procedures and the manager's directions.

3. The manager's decision largely depends on the specific situation and should not be governed by clear strict rules and procedures.

4. The manager's current decision on hiring should be consistently applied in the future.

Item (1) is used to construct the procedure index, item (2) is used to construct the centralisation index, item (3) is used to construct the rules index, and item (4) is used to construct the consistency index.

These four indices are constructed using inverse-covariance weighted (precision-weighted) of the items and normalised to a distribution with a mean of zero and a variance of one.

Enumerator info: Enumerator, please go through this twice. First for scenario 1, then scenario 2.				
Rules Survey - Introduction				
As the final part of this survey we would like to ask experienced Ethiopian managers like you about their philosophy to managing. To do so, we would like to present to you some artificial scenarios to ask how you would act in these. Clearly, there are no right or wrong answers here. Based on what you tell us we will try to infer some patterns which may be consistent with your behaviour. These may also be interesting for you to examine your own decisions. For each of these scenarios, we would like you to imagine I am a colleague coming up to you, sharing some information about a situation in the firm and asking you for your decision. Given the information I give you, you can also ask me for additional information. The goal is to both understand what information you find important making your decision, and what you actually decide. For example, I may come to you with an employee who has missed a day of work without good reason. I would like to know whether we should fire this employee, but before answering you can ask whether the employee has had a previous and/or whether you think it is a good employee. You may then ask me about either of these facts, and then decide whether you would fire the employee. Do you understand the protocol or have any questions?				
	Scenario 1 and 2 - Current			
s1	Firm ID			|_____|
s2	Scenario		01 = scenario 1 02 = scenario 2	|_____|
s3_1	If Scenario 1: Do you have influence on employee compensation at your firm?		01 = Yes 02 = No	|____|
s3_2	If Scenario 2: Do you have influence on employee hiring at your firm?		01 = Yes 02 = No	|____|
Enumerator; tell the respondent: **Scenario 1:** For our first set of scenarios we will talk about an employee who works in your main operation (e.g. an accountant in an accounting firm, a production worker in a manufacturing firm). **Scenario 2:** For our second set of scenarios we will talk about hiring an employee for an entry-level position in your main operation.				
s5	What job does an employee in your main operation do? [Enumerator; just record the job you are focussing on here if there are multiple]	01 = Construction 02 = Administration and Management 03 = Admin/Clerical/Office 04 = Accountant/ Finance/Purchaser 05 = Nurse/Health/Medicine 06 = Teacher/Tutor 07 = Lawyer 08 = Engineer/Architect 09 =Journalist 10 = Psychologist 11 = Banker 12 = Hotel Work (incl. waiter)	15 = Mechanic 16 = Machine Operator 17 = Businessman 18 = Trader/Sales/Retail 19 = Electrician 20 = Driver 21 = Statistician/ Data Collection 22 = Beauty/Hair/Salon 23 = Cleaner/Housework 24 = Transport/Taxi Work 25 = Cook/bakery 26 = Security/Guard/Soldier 27 = Entertainment/Art 28 = Church/Priest 29 = Other	|___|

		13 = Factory 14 = Woodwork and Metal Work/Carpenter/Crafts		
s5_oth	Specify other	If s5 = 29		|_____|

Enumerator, read out: In order to understand your responses better, we would like to limit the scope of information we share with you to the following set of characteristics.
Enumerator; share the sheet with the set of characteristics with the respondent and go through these. Order will be random on your sheets.

Enumerator, tell the respondent:
Only for scenario 1: Suppose I, as your colleague, tell you an employee is unhappy with their pay and would like a raise [next point pay scale] and I want your decision.
Only for scenario 2: Suppose I, as your colleague, tell you I have a candidate for an entry-level position at your firm for which you are hiring. I would like your judgement of whether you would hire this candidate based on what I tell you.

Both scenarios: I will share some information with you, after which you can ask for as many additional facts from this list as you need to make your decision. We would like you to see these scenarios as entirely unrelated. For each scenario, you should decide whether you would give the pay rise/hire the employee, and what information you would like us to share.

Before we start, we would therefore like to ask you whether there is any information in this list that you would always ask you colleague for, regardless of what other information you already have. Secondly, we would like to ask the whether there are any facts you would never find relevant to your decision and thus never ask your colleague about. Answering honestly here will make the survey more relevant for you.

s6	Is there any information from this list you would always need to know to make your decision, regardless of what other information you have?	Enter the up to four characteristics.	C1 C2 C3 C4 C5 C6 C7	|__| |__| |__| |__|
s7	Is there any information from this list you would never ask to know to make your decision.	Enter up to two characteristics.	C1 C2 C3 C4 C5 C6 C7	|__| |__|

Enumerator: After 7, press the done button and check the details on the pop-up window. If these are correct, press confirm to go to the first scenario. If they are not, change these and press done again before pressing confirm to start the first scenario.

Repeat for 8 scenarios:
- Share initial information
- Ask for additional information wanted

- Record decision
- Add comment
- Press confirm

				1	2	3	4	5	6	7	8
s8	What information would you like to have beyond what I told you [Enumerator; make sure the respondent only learns one characteristic at a time].	01 = Yes 02 = No Only enter if 01	C1	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|
			C2	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|
			C3	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|
			C4	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|
			C5	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|
			C6	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|
			C7	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|
s9	What decision does the respondent make?	01 = Yes 02 = No		|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|
s10	Add any comments.		1	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|

Enumerator: Ask about the specific scenarios given by the program once these pop up. Here the respondent is not allowed to ask about additional information.

	Number		1	2	3
S9_ext Decision	What decision does the respondent make?	01 = Yes 02 = No	|_____|	|_____|	|_____|
S10_ext Comment	Add any comments.		|_____|	|_____|	|_____|

Enumerator; when next window pops up:

We are now done with the first part of this scenario. Based on this, we have inferred some rules that other managers may use to make the same decisions as you in the scenarios we asked about. We would like to ask whether you think these patterns are consistent the decision making of your firm.

These rules are of the form: If I know a certain set of characteristics, I do not need any additional information. Going back to our first example before these vignettes a rule may look something like this: If an employee misses a day of work, has a previous warning for this behaviour and is bad at your job, no other characteristic of the employee will make you change your mind.

The rules we suggest here may apply to your decision making with what you told us so far, but because we have not asked about all combinations of the characteristics some may not apply to you.

Enumerator: The first rule will come up on the new window. Ask about this rule and the follow-up questions if the rule applies. Then press OK to go to the next rule. This part is done when the window closes.

			1	2	3	4	5	6	7	8	9	10
s11	Does this rule apply	01 = yes 02 = No [Press confirm to skip to next rule]	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|
s12	What type of	01 = Formal	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|

	rule is this? Is this a formal, informal or personal rule	02 = Informal 03 = Personal Note: **Formal rule**: Written down in organisational documents **Informal rule**: Understood to apply by those in the organisation but not formalised **Personal rule**: Applied by this manager										
s13	Would your inferiors know this rule applies in the firm without asking you or their manager?	01 = Yes 02 = No	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|
s14	Would you always be willing to apply this rule without asking your superior?	01 = Yes 02 = No	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|	|_|

Enumerator info: Enumerator, please go through this twice. First for scenario 1, then scenario 2.				
Rules Survey - Training				
Enumerator, read out to respondent: As the final part of this survey we would like to ask experienced Ethiopian managers like you about their philosophy to managing. To do so, we would like to present to you some artificial scenarios to ask how you would act in these. Clearly, there are no right or wrong answers here. For each of these scenarios, we would like you to imagine I am a colleague coming up to you, sharing information about a situation in the firm and asking you for your decision. The goal is to understand how you would decide in these various scenarios. As an example, I may come to you asking you about an employee who has recently started at the firm who is quite good at their job, but whom has missed a day of work without good reason. I would like to know whether we should fire this employee. Obviously this is somewhat artificial, but we hope to understand your philosophy to these decisions. Do you understand the protocol or have any questions?`				
s1	Firm ID			|____|
s2	Scenario		01 = sc1 02 = sc2	|____|
s3_1	If Scenario 1: Do you have influence on employee compensation at your firm? [if no; skip]		01 = Yes 02 = No	|____|
s3_2	If Scenario 2: Do you have influence on employee hiring at your firm?		01 = Yes 02 = No	|____|
Enumerator; tell the respondent: **Scenario 1:** For our first set of scenarios we will talk about an employee who works in your main operation (e.g. an accountant in an accounting firm, a production worker in a manufacturing firm). **Scenario 2:** For our second set of scenarios we will talk about hiring an employee for an entry-level position in your main operation.				
s5	What job does an employee in your main operation do? [Enumerator; just record the job you are focussing on here if there	01 = Construction 02 = Administration and Management 03 = Admin/Clerical/Office 04 = Accountant/ Finance/Purchaser 05 = Nurse/Health/Medicine	15 = Mechanic 16 = Machine Operator 17 = Businessman 18 = Trader/Sales/Retail 19 = Electrician 20 = Driver 21 = Statistician/ Data	|___|

	are multiple]	06 = Teacher/Tutor 07 = Lawyer 08 = Engineer/Architect 09 =Journalist 10 = Psychologist 11 = Banker 12 = Hotel Work (incl. waiter) 13 = Factory 14 = Woodwork and Metal Work/Carpenter/Crafts	Collection 22 = Beauty/Hair/Salon 23 = Cleaner/Housework 24 = Transport/Taxi Work 25 = Cook/bakery 26 = Security/Guard/Soldier 27 = Entertainment/Art 28 = Church/Priest 29 = Other	

Enumerator, read out: In order to understand your responses better, we would like to limit the scope of information we share with you to the following set of characteristics.
Enumerator; share the sheet with the set of characteristics with the respondent and go through these. Order will be random on your sheets.

Enumerator, tell the respondent:
Scenario 1: Suppose I, as your colleague, tell you an employee is unhappy with their pay and would like a raise [next point pay scale] and I want your decision.
Scenario 2: Suppose I, as your colleague, tell you I have a candidate for an entry-level position at your firm for which you are hiring. I would like your judgement of whether you would hire this candidate based on what I tell you.
Both scenarios: I will share the information for a number of similar cases with you in which an employee asks for a pay rise. We would like you to see these cases as entirely unrelated. For each case, you only need to decide whether you would give the pay rise/hire the employee.

		1	2	3	4	5	6
s9	Decision 01 = Yes 02 = No	|____|	|_____|	|_____|	|_____|	|_____|	|_____|
s10	Comment	|_____|	|_____|	|_____|	|_____|	|_____|	|_____|

** Characteristics for Mail Merge

Scenario 1: Giving an employee a pay rise

C1: Performance compared to other employees with similar jobs (1=above average, 0=below average)

C2: Your colleague personally likes the employee (1=yes, 0 = no)

C3: The employee could easily leave for another firm (1=yes, 0 = no)

C4: The employee has had a pay rise in the past year (1=yes, 0 = no)

C5: Employee's pay is above average at the firm (1=yes, 0 = no)

C6: The employee has worked at the firm for more than two years (1=yes, 0 = no)

C7: The employee's gender (1=male, 0=female)

Scenario 2: Hiring a new employee

C1: The candidate showed enthusiasm for the job (1=yes, 0 = no)

C2: The candidate performed well in the interview (1=yes, 0 = no)

C3: The candidate speaks English (1=yes, 0 = no)

C4: The candidate fits with your firm’s culture (1=yes, 0 = no)

C5: The candidate has relevant experience at another firm (1=yes, 0 = no)

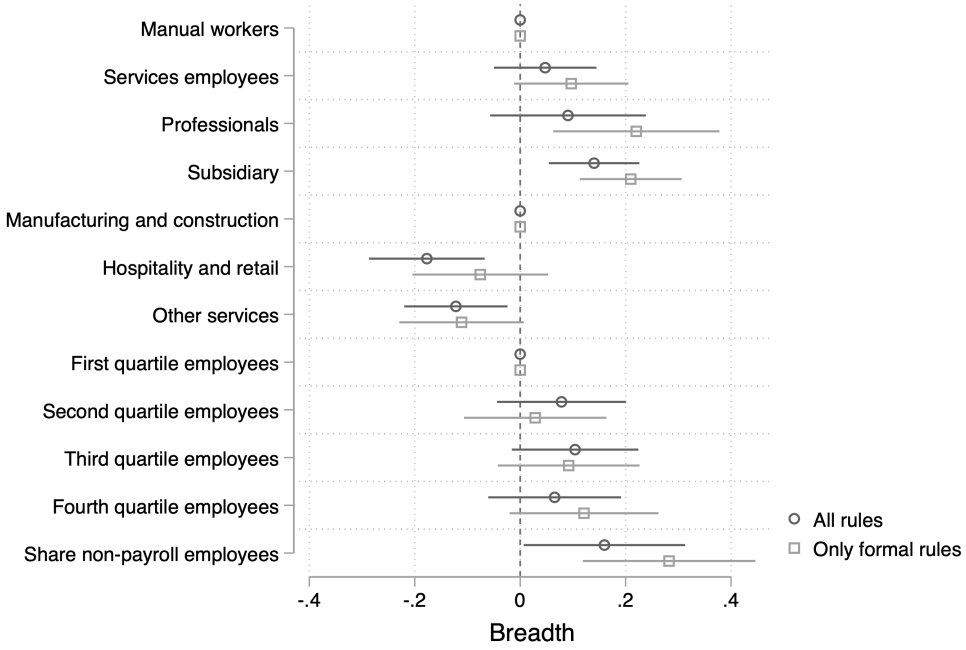
C6: The candidate has a university degree (1=yes, 0 = no)

C7: The candidate's gender (1=male, 0 = female)

2.9.C Additional results

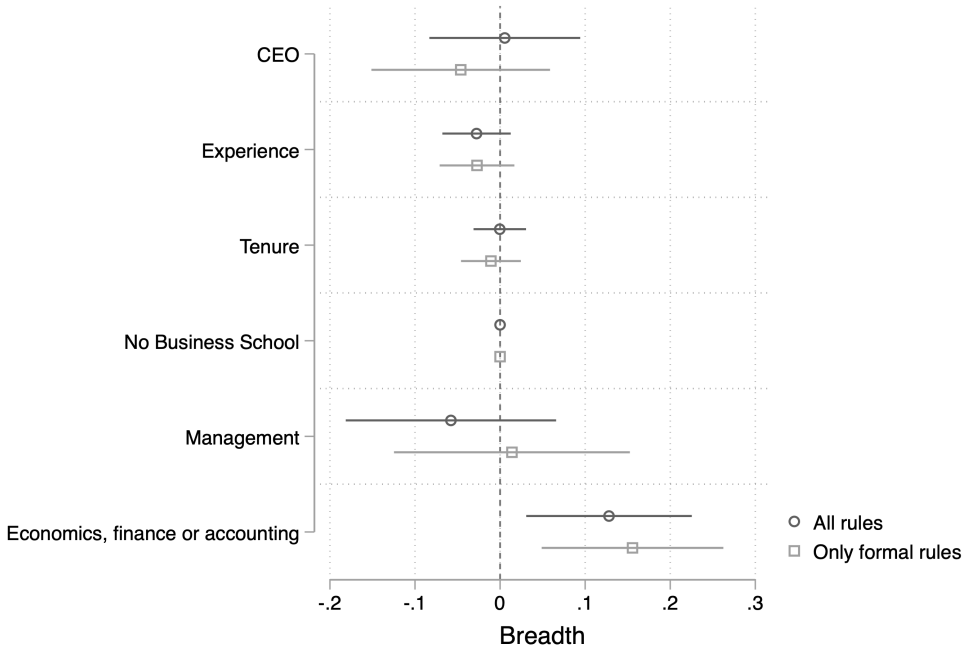
2.9.C.1 Firm and manager observables: Univariate regressions

Figure 2.6: The relationship between firm observables and the breadth of the rulebook.



Notes: This figure reports regression results from univariate regressions of the breadth of the full and formal rulebook on various firm characteristics. These include employee type, categorized into manual workers, service employees, and professionals, with manual workers as the base category; the share of non-payroll employees; a subsidiary dummy variable where 1 denotes that the plant where the interview took place is a subsidiary of a larger entity; industry classification, which includes manufacturing and construction, retail and trade, and other service sectors such as tourism and education, with manufacturing as the base category; and firm size in terms of employment, included by quartile. Error bars represent 95% confidence intervals, with standard errors clustered at the firm level.

Figure 2.7: The relationship between manager characteristics and the breadth of the rulebook.



Notes: This figure reports regression results from both univariate regressions of the breadth of the full and formal rulebook on various respondent characteristics. The CEO dummy indicates whether the respondent is the CEO or equivalent. Experience and tenure are respectively years of experience and years of experience at the firm, measured in units of five years. The management dummy indicates whether the respondent studied management at a business school. The economics dummy indicates if the manager studied economics, finance, or accounting at a business school. Error bars represent 95% confidence intervals.

2.9.C.2 Organisational Rules and Firm Performance

Table 2.18: Organisational Rules and Firm Performance

	Panel A: Sales			
	(1) Sales b/se	(2) Sales b/se	(3) Sales b/se	(4) Sales b/se
Management Score		-19.571		-22.056
		(18.68)		(19.22)
Breadth formal			64.524	70.661*
			(39.25)	(38.97)
Constant	-2.456	-4.046	-11.276	-20.655
	(33.07)	(33.42)	(25.38)	(33.06)
Controls	Yes	Yes	Yes	Yes
N	242	242	242	242
Adj-R2	0.1672	0.1739	0.1849	0.1880
	Panel B: Profit			
	(1) Profit b/se	(2) Profit b/se	(3) Profit b/se	(4) Profit b/se
Management Score		-0.577		-0.768
		(0.80)		(0.91)
Breadth formal			6.538**	6.632**
			(2.78)	(2.85)
Constant	1.017	0.908	0.164	-0.394
	(3.07)	(3.10)	(2.17)	(3.17)
Controls	Yes	Yes	Yes	Yes
N	232	232	232	232
Adj-R2	0.0341	0.0316	0.0697	0.0631
	Panel C: Profit over sales			
	(1) Profit/Sales b/se	(2) Profit/Sales b/se	(3) Profit/Sales b/se	(4) Profit/Sales b/se
Management Score		-0.006		-0.008
		(0.02)		(0.02)
Breadth formal			0.081*	0.083*
			(0.04)	(0.04)
Constant	0.203***	0.202***	0.186***	0.185***
	(0.05)	(0.05)	(0.05)	(0.05)
Controls	Yes	Yes	Yes	Yes
N	230	230	230	230
Adj-R2	0.0418	0.0381	0.0576	0.0546

Notes This table examines how predictive the formal rule breadth is for sales, profit and profit over sales relative to the MOPS management score including the HR question. The calculation of the management score and formal breadth are detailed earlier in this section. All standard errors are clustered at the firm level. The included controls are the number of employees, sector, the type of position, management practices, and whether the firm is a subsidiary. Figure 2.1 details these variables, the number of employees is incorporated as a continuous variable rather than by quartile. Standard errors, clustered at the firm-level, are indicated in braces and significance is indicated as: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.9.D Details of the trained Gaussian Process

This section provides a more detailed description of the implemented Gaussian Process, the estimated parameters based on training data and their interpretation.

The linear mean function is represented as:

$$m(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} + \alpha$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_7]^\top$ is the vector of characteristics (e.g., performance, tenure, etc.). $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_7]^\top$ are the coefficients representing the contribution of each characteristic to the mean function. α is the intercept term.

The covariance structure of the Gaussian Process is defined by a squared exponential kernel with Automatic Relevance Determination (ARD):

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{7}\frac{(x_{i,d} - x_{j,d})^2}{\ell_d^2}\right)$$

where σ_f^2 is the variance parameter, controlling the overall vertical variation, $x_{i,d}$ and $x_{j,d}$ are the d-th characteristics of the inputs $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. ℓ_d is the length-scale parameter for the d-th characteristic, controlling how much that characteristic influences the similarity between two input points (i.e., how sensitive the covariance is to changes in that particular characteristic).

The Gaussian Process prior over the latent function $f(\mathbf{x})$ is:

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}_i, \mathbf{x}_j)\right)$$

where $m(\mathbf{x})$ is the linear mean function and $k(\mathbf{x}_i, \mathbf{x}_j)$ is the squared exponential kernel with ARD.

The logistic link function used to map the latent output into the probability space is:

$$p(y = 1 \mid \mathbf{x}) = \sigma(f(\mathbf{x})) = \frac{1}{1 + \exp(-f(\mathbf{x}))}$$

where $f(\mathbf{x})$ is the latent output from the Gaussian Process and $\sigma(\cdot)$ is the logistic

(sigmoid) function, which transforms the latent variable into the range $(0, 1)$.

These equations describe the components of your Gaussian Process model, including the linear mean function, the squared exponential kernel with ARD, and the logistic link function for probabilistic interpretation.

Table 2.19: Parameters for Gaussian Processes: Pay Rise and Hiring Scenarios

Characteristic	Mean parameter	**Value**	Kernel parameter	**Value**
Performance compared to others	β_1	0.1404	ℓ_1	0.5132
Colleague likes the employee	β_2	0.2101	ℓ_2	0.5362
Employee could leave for another firm	β_3	0.2662	ℓ_3	0.9306
Had a pay rise in the past year	β_4	0.1667	ℓ_4	0.0451
Pay is above average at firm	β_5	0.1038	ℓ_5	0.5377
Worked at firm $>$ 2 years	β_6	0.1821	ℓ_6	1.3443
Employee's gender, male=1	β_7	0.1623	ℓ_7	0.2442
	α	-0.8449	σ_f^2	0.3197
Characteristic	Mean parameter	**Value**	Kernel parameter	**Value**
Candidate showed enthusiasm	β_1	0.2092	ℓ_1	0.6234
Performed well in interview	β_2	0.2943	ℓ_2	0.0914
Speaks English	β_3	0.4317	ℓ_3	0.8466
Fits firm's culture	β_4	0.3491	ℓ_4	0.5290
Relevant experience	β_5	0.2273	ℓ_5	1.6567
Has university degree	β_6	0.2545	ℓ_6	0.3127
Candidate's gender, male=1	β_7	0.2252	ℓ_7	0.3207
	α	-1.2087	σ_f^2	0.1289

This table shows that the probability of a positive response is increasing in each of the characteristics. This is as expected for all the characteristics, except for gender where the relationship with decisions was ex-ante unclear. In future analysis, this could be implemented in a regression framework to better understand the results.

In the pay-rise scenario, the probability of the decision being one varies from 30% when $X = 0$ for all characteristics to 60% when $X = 1$ for all characteristics. For the hiring scenario, this range is broader and goes from 23% to 69%.

Lights, Camera, Transaction: Assessing Management Styles through Studio Vignettes

Girum Abebe
Marcel Fafchamps
Michael Koelle
Simon Quinn
Tom Schwantje

Abstract

We introduce a new method for measuring managerial traits of young professionals: using management vignettes in a video studio. This method – analysed through the lens of a Bayesian hierarchical model – allows us to identify four distinct managerial archetypes (which we term 'authoritative', 'affiliative', 'coercive' and 'timid'). We find that past labour market exposure (including exposure induced through a previous field experiment) correlates strongly with the propensity to act as an authoritative manager. We then use the videos to run an incentivised experiment with firm managers, to elicit preferences over young professionals. Strikingly, we find that firms consistently prefer authoritative-type managers for entry-level managerial positions. Empirically, our results highlight an underexplored mechanism for labour market exclusion among young professionals. Methodologically, we demonstrate the value of controlled vignette scenarios for assessing managerial traits. Our findings underscore the importance of managerial training in shaping labor market outcomes, and offer new avenues for studying the development of managerial talent.

3.1 Introduction

The late twentieth century witnessed a marked expansion of formal education across many urban low-income settings (Alesina, Hohmann, Michalopoulos, and Papaioannou, 2021; Asher, Novosad, and Rafkin, 2022). As a result, urban labour markets in low- and middle-income countries have undergone a striking demographic shift – whereby a substantial share of young job seekers possess significantly higher levels of formal education than their parents. This educational mobility has important implications for understanding barriers to labour market inclusion – particularly among those aspiring to professional jobs that, for many, differ markedly from the roles held by their parents. It is well recognised that formal schooling contributes to the accumulation of human capital, as well as to the development of non-cognitive skills and socialisation (Bowles and Gintis, 1977, 2002; Heckman and Rubinstein, 2001). It remains unclear, however, how family background affects individuals' ability to socialise into norms of professional jobs. This is an issue of first-order relevance to understanding the large and growing literature on active labour market programs in low-income settings (Caria, Orkin, and Bedoya, 2023) – and, more generally, for understanding the allocation of professional talent in such economies (Feng, Lagakos, and Rauch, 2024). However, it is an issue that has received almost no attention in the empirical literature.

To make progress on this issue, we design and implement a novel style of controlled experiment to measure managerial traits at the individual level. We use this measure to test whether professional traits constitute a barrier to labour market inclusion for educated youth. Specifically, we work with a large sample of young professionals, chosen for their interest in business and entrepreneurship – many of whom are not yet working in managerial roles. We invite these young professionals to a video studio, where we explain that they will each watch a series of actors portraying different managerial scenarios. We ask the young professionals to record their responses – as if they are in an entry-level managerial position in a hypothetical firm. With the agreement of those young professionals, we then show these videos to human resources managers in established

medium and large manufacturing firms – using an incentive-compatible mechanism to elicit the firms' assessments of the suitability of the young professionals both for wage jobs as entry-level managers and for running their own firms. By designing this controlled experimental setting, we are able to pose the same set of managerial challenges to every respondent in this sample – avoiding the endogenous assignment of managerial tasks to managers that is otherwise inherent in working with observational data on managerial decisions.

Central to our approach is the notion that managerial traits can be quite impressionistic; different managerial styles are often best understood as complex bundles of behaviours, rather than being straightforward to observe and record (Benson and Shaw, 2025; Goleman, 2000). For this reason, we deliberately chose a diverse range of managerial problems (rather than, for example, repeating a single class of scenarios), and we encouraged open-ended answers (rather than, for example, forcing our respondents to choose from a small menu of available actions). Similarly, we then encode respondents' answers in a high-dimensional space (specifically, we measure responses through the combination of managerial action, justification, tone, and source of authority). To analyse this data, we build and estimate a bespoke Bayesian hierarchical model. This has the flavour of Latent Dirichlet Allocation ('LDA') models (Bandiera et al., 2020; Blei, Ng, and Jordan, 2003; Griffiths, 2004) and – like LDA models – allows us to characterise heterogeneity in managerial styles across a high-dimensional response space.

We have three key results. First, our method – that is, the combination of a studio exercise with a Bayesian hierarchical model – succeeds in identifying meaningful and substantial differences in managerial traits across young professionals. Specifically, we describe four latent 'pure types' of managerial style. After estimation – and following Goleman's (2000) classic work in managerial studies – we label these types as '*authoritative*', '*affiliative*', '*coercive*' and '*timid*'. These types differ quite radically in their conceptualisation of the role of a junior manager. For example, coercive managers seem to view their role in terms of implementing their personal objectives, and do so relying upon their personal authority. Affiliative managers seek shared ground – often yield-

ing to the other side in their responses, and emphasising shared interests as they do so. Authoritative-type managers see themselves as implementing rules set by more senior managers: they are much more likely than other types to rely upon formal policy, and they emphasise the firm's interests in doing so.

Second, we find a clear preference across firms for authoritative-type managers over managers of the other three types. We designed our experiment with a view that some firms may (for example) have a clear preference for authoritative-type managers, whereas (for example) others may prefer more affiliative types (see, in particular, Bandiera et al. (2020)). It is striking that we do not find this kind of heterogeneity: while we do detect some heterogeneity in strength and ranking of different styles, firms seem largely united in their view that the authoritative style of management is best – both for entry-level managerial roles and for self-employment. After obtaining our main Bayesian estimates, we returned to the field to run semi-structured interviews with firm employees; these interviews show that employees share their managers' preferences for the authoritative type.

Third, young professionals' managerial behaviour is strongly related to their past exposure to the labour market. In particular, we find that authoritative-type respondents are more likely than the other types to have been self-employed, to earn higher wages (conditional on wage employment), and to report higher reservation wages and reservation profits. Strikingly, they are also significantly more likely to have been in the treatment group in a previous field experiment that randomly increased some respondents' exposure to medium and large firms (see our related work, Abebe, Fafchamps, Koelle, and Quinn (2024)); this implies that the relationship between labour market exposure and managerial style is causal. We find that this experimental effect is driven entirely by individuals whose parents did not finish primary school. This implies that the impact of early-career labour market exposure is particularly valuable for those who are traditionally likely to be excluded from labour market opportunities – and that lack of 'professional socialisation' is likely to be a key mechanism by which labour market inequality can persist across generations.

Together, these results contribute to three bodies of literature. First, our results complement the recent literature studying labour market exclusion among young professionals. Our results show that individuals with greater labour market exposure are more likely to exhibit the authoritative style of management – and, conversely, that the authoritative style of management is in higher demand among prospective employers. Together, these two results imply that embodied managerial traits may act as a key mechanism by which some young professionals enjoy sustained labour market success while others face exclusion. This supports a key insight from our earlier work (Abebe et al., 2024) – and, indeed, our results here help to interpret the mechanisms identified in that earlier paper. More generally, our results highlight the importance of non-cognitive skills for employment in urban low-income settings (Bassi and Nansamba, 2022); indeed, our result on the relevance of parental education suggests that embodied managerial traits may act as a form of 'cultural capital' (Bourdieu, 1986), by which existing labour market inequalities are reinforced (Barrios-Fernández, Neilson, and Zimmerman, 2024; Falk, Kosse, and Pinger, 2020; Shukla, 2025; Zimmerman, 2019).

Second, our findings add to our understanding of managerial traits. The past two decades have witnessed a substantial expansion in economists' interest in the role of management. This has been driven by a new empirical focus on the measurement of management practices, spearheaded by the World Management Survey (Scur, Sadun, Van Reenen, Lemos, and Bloom, 2021). A related literature has highlighted the critical role of managers themselves, using comparisons both across firms (Bertrand and Schoar, 2003; Bloom et al., 2019) and within firms (Hoffman and Tadelis, 2021; Lazear, Shaw, and Stanton, 2015; Metcalfe, Sollaci, and Syverson, 2023). Managerial *traits* embodied in *individual managers* have been have been the focus of a smaller literature in economics (see, in particular, Malmendier, Tate, and Yan (2011), Kaplan, Klebanov, and Sorensen (2012), Bandiera, Guiso, Prat, and Sadun (2015), Benmelech and Frydman (2015), Bandiera et al. (2020) and López-Peña, Mozumder, Rabbani, and Woodruff (2025)). This literature has studied decisions taken by managers, in their managerial capacity; that is, samples of respondents who (i) have already been selected by firms for

their managerial abilities, and (ii) whose managerial challenges are determined by the particular managerial contexts in which they find themselves (Weidmann, Vecci, Said, Deming, and Bhalotra, 2024). In contrast, by measuring managerial traits through a controlled studio environment, we are able to shed new light on the role of such traits as a mechanism for labour market exclusion. That is, we are able to measure managerial traits even among those who are not employed as managers.

On the one hand, our result that managerial traits are related to past labour market exposure is broadly consistent with earlier results showing the relevance of past experiences in shaping managerial style (Benmelech and Frydman, 2015; Malmendier et al., 2011). In contrast, our result on the relative homogeneity of firm preferences provides a counterpoint to the seminal earlier work of Bandiera et al. (2020) – who find heterogeneous firm-manager match quality among different kinds of employed senior managers. This presents a novel insight on firm preferences over management: while different firms may have different preferences over the managerial styles of their *senior* leaders, firms seem reasonably homogeneous in their preferences for *entry-level* managers among a broad pool of young professionals.

Third, methodologically, our results contribute to recent literature using innovative and open-ended data collection methods in economics. This includes – in particular – recent literature on the relevance of tone and expression in economic communications (Chang, Dai, Feng, Han, Shi, and Zhang, 2025; Gorodnichenko, Pham, and Talavera, 2023; Handlan and Sheng, 2023), the use of photographs to proxy labour market potential (Guenzel, Kogan, Niessner, and Shue, 2025) and pre-trial detention (Ludwig and Mullainathan, 2023), and on the use of machine learning techniques to understand CEO performance (Borgschulte, Guenzel, Liu, and Malmendier, 2021). More generally, our approach complements recent advances in the use of open-ended survey questions in economics (Haaland, Roth, Stantcheva, and Wohlfart, 2024; Stantcheva, 2021). Separately, our paper illustrates and validates the use of studio vignettes as a credible method for eliciting managerial traits in field settings. This builds on a small literature in economics that has used a different kind of team-based 'management simulations', to test the impacts of

managerial training in the field (Macchiavello, Menzel, Rabbani, and Woodruff, 2020) and to assess managerial potential in the lab (Weidmann and Deming, 2021; Weidmann et al., 2024). More generally, we note that the use of video responses to pre-recorded questions is increasingly common as part of 'assessment centres' for job recruitment exercises and university entrance processes – but has received relatively little empirical attention in economics.

The paper proceeds as follows. In section 3.2, we describe the experimental context and implementation. In section 3.3, we use a Bayesian hierarchical model to characterise heterogeneity in management traits. We go on in that section to describe the characteristics of those types, and then provide a sentence-embedding analysis of different types' vocabulary. In section 3.4, we turn to the firm side – to characterise firm preferences over different managerial approaches (including exploiting an alternative source of exogenous variation – namely, random variation in the gender of the actors viewed). We go on to report a model-informed field survey of workers. Section 3.5 shows the impact of our earlier management placement upon respondents' management styles and firms' assessments of those respondents. Section 3.6 concludes.

3.2 The experiment

3.2.1 Vignettes in the studio

Our experiment is designed to measure management traits among young professionals across a series of realistic management scenarios. To do this, we used a studio setting, in which respondents participated in a role-play scenario. Specifically, we ran a series of separate vignettes; in each vignette, the respondent watched a video of a paid actor, who played the role of a counter-party in a managerial problem. For each vignette, we played the video (explaining that it showed an actor), and then asked the respondent to provide a short response – as if she or he were in a managerial role, responding to the character in the video. We explained to the respondents that their video recordings would be played to human resource managers in successful Ethiopian firms. We recorded

each vignette using both a male actor and a female actor (with different actors for the different vignettes); we randomly varied whether respondents viewed the male or female actor, and we exploit this random variation in section 3.4.2.[5]

Specifically, we set each respondent five different scenarios:[6]

1. *Line management of an employee:* The actor is an accounting clerk who has been absent for three days without warning. (S)he shows up for work on the fourth day, explaining that (s)he was unwell.

2. *Negotiating with a supplier:* The actor plays a supplier, who explains that (s)he cannot fulfil an order according to specifications, because of problems sourcing input materials. (S)he offers to supply a replacement of inferior quality instead. Respondents are told specifically that the firm they represent is known for producing and selling the highest quality products in the industry.

3. *Negotiating a pay rise:* The actor is a production worker, who comes to ask for a pay rise. The worker argues that (i) within her unit, (s)he has been the most productive worker during the last three months, and one of the few people who exceeded the personal productivity targets; and (ii) (s)he has been with the company for ten years, and even her mother used to be an employee of the company. Respondents are told that the firm they represent has no plan to increase any salaries this year.

4. *Negotiating an adjustment with the bank:* The actor is a bank manager, who calls to remind the firm about an unpaid loan installment (explaining that failure to pay increases interest payments and reduces the firm's credit rating). Respondents are told that the firm they represent will not have sufficient funds to pay the bank for another two weeks.

5. *Negotiating with a client:* The actor is a client, who has not paid what you have invoiced, in spite of one reminder. The client comes to place a new order.

[5] Amharic marks second-person agreement in both pronouns and verb conjugation – using masculine or feminine forms for gendered addressees, plus a gender-neutral honorific. In nine of our ten videos, the speaker uses the honorific form; in the tenth video, the speaker begins with a gendered pronoun and corresponding verb form before switching to the honorific.

[6] The full scenario scripts are provided in Appendix 3.7.A.

3.2.2 Incentivised elicitation of firm preferences

In the second part of our experiment, we played the studio recordings to human resources managers from successful Ethiopian firms.[7] For each vignette, the human resources manager watched three responses, and was asked to rank these separately based on (i) the manager's assessment of the respondents' suitability as entry-level managers at their firm, and (ii) the manager's assessment of the respondents' suitability to run their own small business.

Specifically, each human resources manager was assigned to assess video recordings from three different respondents, for each of the five vignette scenarios. Each triplet of young professionals was assessed by at least two different managers for each vignette. To elicit revealed preferences comparing different candidates, we implemented two complementary direct elicitation mechanisms: one to elicit the perceived suitability of candidates for a managerial position in the firm; the other to elicit candidates' perceived suitability to run their own business. We incentivised the assessment of suitability for managerial positions through the prospect of receiving respondents' contact details (with respondents' permission), and we incentivised the assessment of suitability to run a business through the prospect of the respondent being invited to a business plan competition. They were designed to be straightforward and to make truthful ranking an 'obviously dominant' strategy for respondents.[8] Separately, we asked each human resources manager directly to assess the individual respondents (again, both for their suitability as an entry level manager, and as an entrepreneur), using a series of stated-preference questions. We explain both the revealed-preference and stated-preference methods in detail in Appendix 3.7.B.

[7] We ran this assessment with the senior member of the firm who was responsible for hiring. We refer to this person here as 'the human resources manager' – though this was not always the actual title or job description.

[8] Both mechanisms closely resembled the 'OSP-RSD' ranking mechanism described by Li (2017).

3.2.3 Implementing our experiment

We conduct our experiment with 982 young professionals, who previously participated in an experiment in which a random subset was assigned to a management placement (see Abebe et al. (2024)).[9] The respondents are primarily male (78%), highly educated (78% have a university degree), and relatively young (with an average age of 31). Most of the young professionals are in wage employment (68%) or self-employment (17%). Of those in wage employment, 98% work in a professional position, and 18% in a managerial position. Conditional on wage employment and self-employment, respectively, the median wage and profit are both 6000 ETB (although self-employment income is more disperse across respondents). This is more than twice the average household expenditure of 2500 ETB in Addis Ababa.[10]

Despite our respondents' high level of education, there is significant variation in their socio-economic backgrounds. Only 11% of our respondents have a parent who completed a university degree; even more striking, 48% of respondents report that neither parent completed primary school.[11] These patterns emphasise that these young professionals are participating in a labour market that differs markedly from the one their parents faced. Appendix Table 3.36 summarises.

The sample of firms was drawn from a set of 713 established Ethiopian firms involved in a previous experiment (Abebe et al., 2024), supplemented by additional firms to account for attrition. These are medium-sized to large firms operating across a range of industries, in Addis Ababa and in the neighbouring towns of Bishoftu and Adama. Appendix Table 3.38 describes the sample included in the project. We conducted interviews with a sample of 576 firms, where we interview a manager responsible for human resource decisions in the firm. These firms have a median of 58 employees (mean = 323) and employ a median of five managers (mean = 12). The interviewed managers work

[9] We show robustness to attrition in section 3.5.

[10] See the Ethiopian Socioeconomic Survey (Ethiopian Statistical Service and World Bank, 2023).

[11] Alesina et al. (2021) show that, for urban Ethiopia generally, the 'upward intergenerational mobility' – that is, the probability of completing primary school conditional upon neither parent having completed primary school – is about 55% (see their Appendix Table E.II). For a broader sample of 28 countries, Ouedraogo and Syrichas (2021) find that individuals born in 2000 and residing in urban areas exhibit, on average, upward intergenerational mobility of 78%.

primarily in human resources (40%) and administration (34%). They have been in their current position for, on average, eight years. 80% have a university degree and 72% report having formal management training. Appendix Table 3.37 summarises.

3.3 Management styles in the studio

Our first experimental objective is to characterise the main management styles among young Ethiopian professionals. To do this, we build a bespoke Bayesian Hierarchical Model; this allows us to identify distinct clusters of managerial behaviour. To do this, we first employed two enumerators to watch all of the videos, and to encode each response along the following four dimensions:[12]

1. *Action*: Did the respondent agree or disagree with the actor's request? (For example, in the third scenario, did the respondent agree to the employee's request for a pay rise?)

2. *Authority*: What source of authority did the respondent rely upon? Specifically, did the respondent rely upon (a) their formal authority as a manager, (b) their seniority or personal authority, (c) higher principles, such as an appeal to the appropriateness of the action, or (d) did the respondent not rely upon any source of authority.

3. *Justification*: How did the respondent justify her or his action? Specifically, did the respondent (a) provide no justification, (b) emphasise the interests of the firm they represent, (c) emphasise the interest of the other party, (d) emphasise the shared interest of the firm and the other party, or (e) emphasise their personal interest?

4. *Tone*: What tone did the respondent use? Specifically, was it (a) calm/assured, (b) assertive, or (c) aggressive?

[12] Appendix 3.7.C provides a detailed description.

3.3.1 Management traits among Ethiopian young professionals

We begin, in Figure 3.1, by describing the average behaviour across our five vignettes. Specifically, Figure 3.1 shows one row for each vignette; across that row, we characterise the average behaviour across each dimension of management (as discussed, these are action, authority, justification and tone). We disaggregate the data by the two enumerators. We draw two conclusions from the figure. First, we find substantial heterogeneity *across* different vignettes – and this is reflected across the bundle of different managerial dimensions. For example, when dealing with an employee seeking a pay rise (vignette 3), most respondents refuse the request; to do so, they rely heavily on formal authority and on personal authority, and their tone can be relatively assertive. In contrast, when dealing with their bank (vignette 4), respondents are much more likely to accede to the request, to place almost no reliance upon formal policy, and generally to use a less assertive tone. Similarly, when dealing with an employee seeking a pay rise (vignette 3), respondents frequently invoke their shared interest, but this rarely happens when dealing with a supplier failing to deliver a high-quality good (vignette 2).

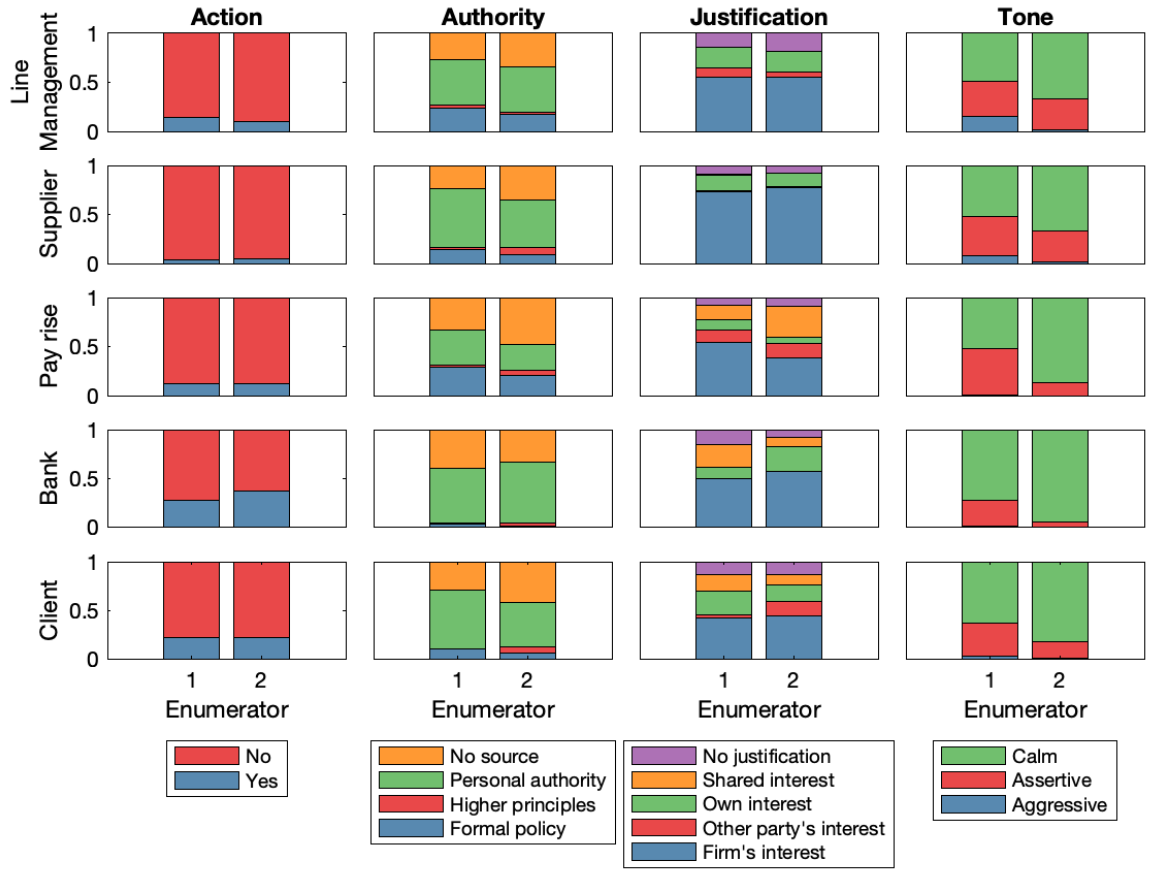
Second, there is substantial heterogeneity *within* different vignettes, across respondents. This is evident from the disaggregation across the separate enumerators. Despite assessing the videos independently, the enumerators describe the responses in broadly similar terms – implying that the observed differences in action, authority, justification and tone are driven by heterogeneous approaches by different young professionals, rather than by differences in enumerators' perceptions.

3.3.2 A Bayesian model of management styles

To extend this analysis, we now build a bespoke Bayesian Hierarchical Model. This model serves two related purposes. First, it allows us to cluster the different dimensions of managerial responses into different managerial styles. Second, the resulting classification will be a key input for understanding firm preferences in the following section.[13]

[13] For this section, we only use the data from the enumerator encoding of the vignettes to estimate our model.

Figure 3.1: The distribution of behaviours by vignette



Notes This figure reports the distribution of the encoded behaviours by vignette and by enumerator. The bars show the probability that an enumerator encodes a specific behaviour for each dimension by vignette.

To do this, we specify a generative model in which each young professional – when responding to a given vignette – chooses a combination of action, authority, justification and tone. The model characterises 'pure types' that represent archetypal managerial styles; each young professional is then represented as a convex combination of those archetypes. A Bayesian Hierarchical framework is well suited for this task due to its ability to handle high-dimensional categorical data and to uncover latent themes. Very broadly, this is similar to several recent papers on heterogeneity in management styles (e.g. Bandiera et al., 2020).[14]

Specifically, we observe individuals $i \in \{1, \ldots, N\}$ performing on vignettes $v \in \{1, \ldots, 5\}$. For each individual assessment of a vignette, we have two enumerators, $e \in \{1, 2\}$. Each enumerator records a set of 'attributes' of the response (action, authority, justification and tone): $a \in \{1, \ldots, 4\}$. Each attribute a has $J(a)$ possible categorical responses, $y^a_{ive} \in \{1, \ldots, J(a)\}$.[15] For each vignette, each respondent draws their behaviour from one of K pure types, z_{iv}. The total probability of the model is as follows:

$$P(W, Z, \boldsymbol{\theta}, \boldsymbol{\phi}, \zeta, \eta) = \underbrace{\prod_{j=1}^{K} P(\boldsymbol{\phi_j}; \zeta)}_{\text{Dirichlet parameters}} \cdot \underbrace{\prod_{i=1}^{N} P(\boldsymbol{\theta_i}; \eta)}_{\text{Types}} \cdot \prod_{v=1}^{5} \sum_{k=1}^{K} \left(\underbrace{P(z_{iv} = k | \boldsymbol{\theta_i})}_{\text{Type assignment}} \cdot \underbrace{\prod_{e=1}^{2} \prod_{a=1}^{4} P(y^a_{ive} | \boldsymbol{\phi}_{ka}, \boldsymbol{\psi}_{av}, \boldsymbol{\chi}_{ae})}_{\text{Studio behaviour}} \right), \tag{3.6}$$

where $P(\boldsymbol{\phi}_j; \zeta)$ and $P(\boldsymbol{\theta}_i; \eta)$ follow a Dirichlet distribution, $P(z_{iv} | \theta_i)$ follows a categorical distribution and $P(y^a_{ive} | \boldsymbol{\phi}_{ka}, \boldsymbol{\psi}_{av}, \boldsymbol{\chi}_{ae})$ follows a Multinomial Logit distribution.

In this model, $\boldsymbol{\phi}_{ka}$ characterises the behaviour of the pure types and $\boldsymbol{\theta}_i$ characterises an individual as a convex combination of those archetypes. z_{iv} is an individual's type for a specific vignette. $\boldsymbol{\psi}_{av}$ is a vignette fixed effect that allows average behaviour to vary across vignettes, and $\boldsymbol{\chi}_{ae}$ is an enumerator fixed effect that allows average behaviour to vary across enumerators. We estimate using a Hamiltonian Monte Carlo algorithm in

[14] Our model has some analogy to the class of Latent Dirichlet Allocation ('LDA') models. In the language of LDA models, we could think of each respondent-vignette pair as comprising a 'document', each document comprising four 'words', and each word being drawn respectively from one of four different dictionaries. Our inclusion of enumerator and vignette effects is a further variation on the typical LDA setup, akin to what Biderman, Blei, Cai, Ciccarone, Feder, and Prat (ress) term a 'fixed effects topic model'.

[15] Specifically, $J(1) = 2$ for the action; $J(2) = 4$ for source of authority, $J(3) = 5$ for justification and $J(4) = 3$ for tone.

Stan (Stan Development Team, 2024); we discuss the parameterisation and estimation of this model in more detail in Appendix 3.7.E; standard MCMC diagnostics indicate excellent convergence.

3.3.3 Types of management

We estimate our model allowing for four 'pure types' of management. (In Appendix 3.7.F, we consider alternative versions with two, three and five types and show that the general conclusions remain very similar.) Figure 3.2 shows the four estimated 'pure type' management styles and their behaviour in our five vignettes:

1. Type 1 implements the rules set by the firm. Specifically, they refuse the requests almost all of the time; in doing so, they are much more likely than the other types to rely upon formal policy – and much more likely to emphasise the firm's interests in doing so. They are more likely than the other three types to use an assertive tone; indeed, at times, they can even be aggressive.

2. Type 2 seeks shared ground. Strikingly, the affiliative type is much more likely than the other three types to concede to their counterparty in the hypothetical discussion – notwithstanding that many of the vignette prompts suggested that such concessions were not in the interests of the hypothetical firm. The affiliative type tends to rely upon personal authority (or cites no specific authority), and is much more likely than the other three types to emphasise shared interests. Their tone is generally calm, but can be assertive.

3. Type 3 implements their personal objectives. They refuse the requests almost all of the time; in doing so, they are overwhelmingly likely to rely on personal authority. As justification, they tend to emphasise the firm's interest – but are much more likely than the other three types to justify their actions by reference to their own personal interest. Similar to the authoritative type, they tend to be assertive, or even aggressive.

4. Type 4 does not provide a clear explanation for their decision. They refuse the requests almost all of the time – and, in doing so, are unlikely to rely on any source of authority, and often provide no specific justification for their decision. Of the four pure styles, timid managers are most likely to be calm in their tone.

In Figure 3.3, we describe the distribution of individuals over types. To do this, we draw a tetrahedron, in which each vertex represents probability one of belonging to a particular type. We learn three things from this figure about our model results. First, the model distributes individuals quite evenly across the tetrahedron, such that any given individual exhibits a combination of different 'pure type' behaviours. Second, very few individuals sit close to the affiliative vertex: this implies that affiliative behaviour, while empirically important, is something that individuals incorporate into their responses – rather than being a style that consistently describes any particular individual. Third, the stacked bar on the right of Figure 3.3 shows the observed vignette responses in terms of the average distribution of each pure type across individuals. We estimate that the most prevalent are type 1 (28.0%) and type 3 (27.9%); this is followed by type 4 (25.2%) and then type 2 (19.0%).

3.3.4 Managerial types and labour market experience

Having estimated the distribution of individuals across types, we now study the labour-market outcomes and trajectories of these individuals, conditional on their managerial traits. Panel A of Table 3.1 reports the average characteristics of individuals by type assignment. The authoritative type is most likely to be male and self-employed and, relative to the other types, is more likely to report an above-median reservation wage and profit. The authoritative type is also most likely to have completed at least a bachelor's degree. Differences among the remaining three types are more subtle, yet several stand out. The affiliative type is most likely to be female; the timid type tends to report a below-median reservation wage; and the coercive type is most likely to have an above-median reservation profit.

Next, we examine how labour-market experience varies across managerial types.

Figure 3.2: 'pure type' management styles amongst Ethiopian young professionals

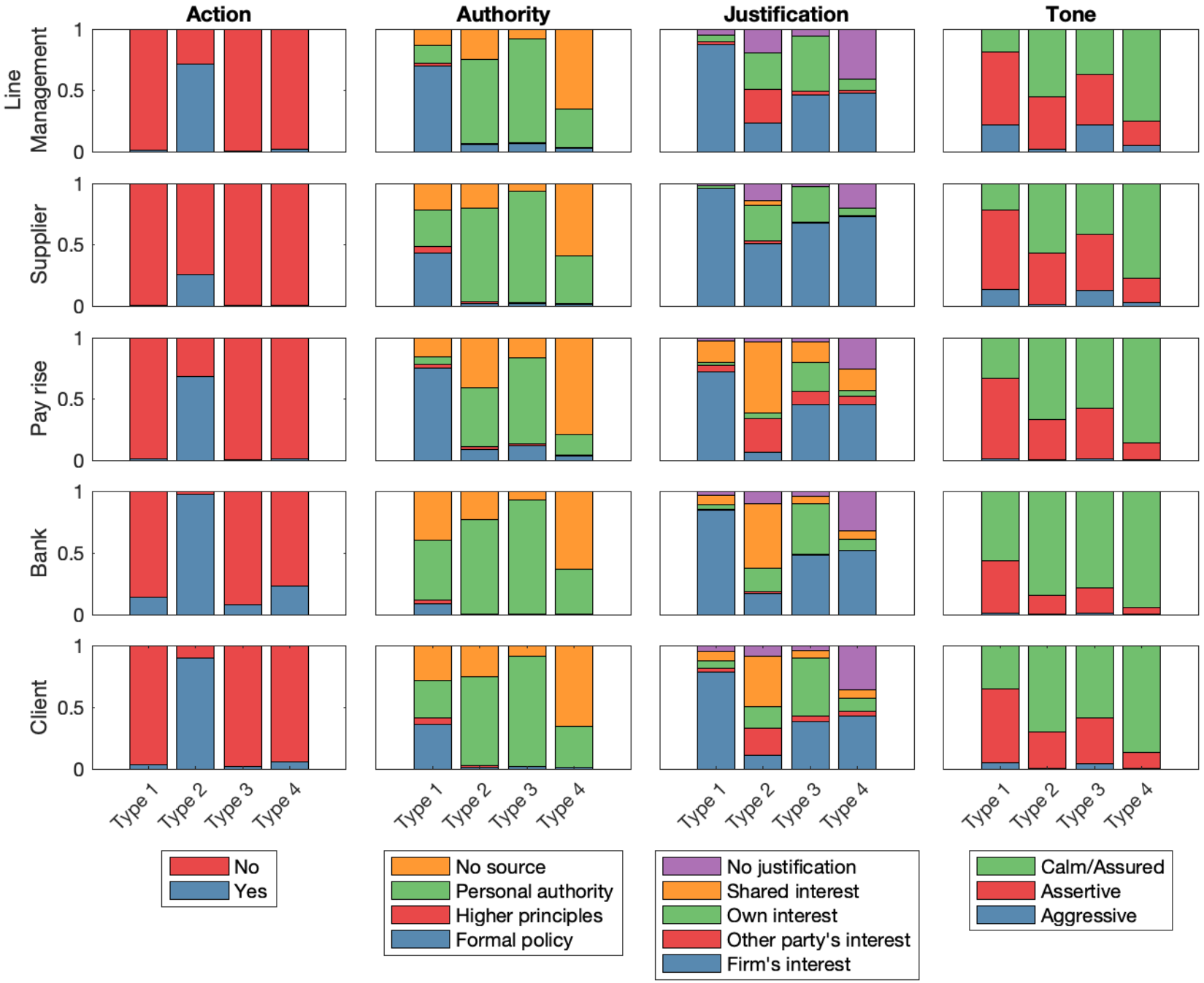
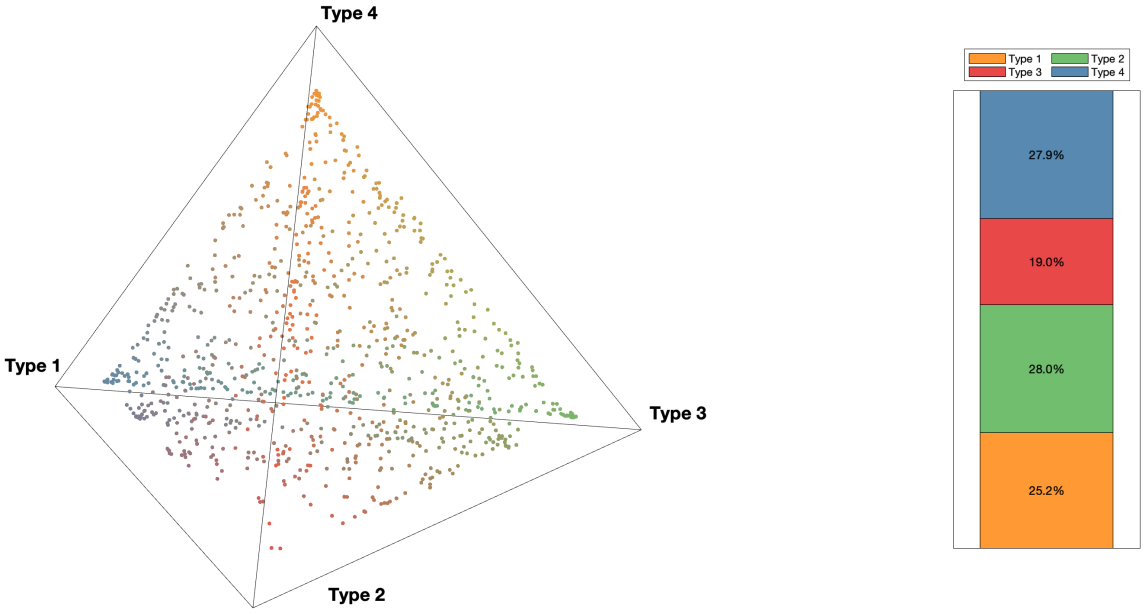


Figure 3.3: Distribution of types across individuals



Approximately 12 months after respondents participated in the studio, we asked them about their labour-market experiences over the previous six years, including their employment status at six-month intervals. Panel B of Table 3.1 summarises these experiences by managerial type. The table indicates that authoritative-type managers spent more years in employment—particularly in permanent positions—and made about 0.2 fewer job transitions on average than other types, reflecting a more stable employment history. Finally, authoritative managers were also most likely to be in a management position when we interviewed them before participating in the studio.

Table 3.1: Characteristics and Types: Summary Statistics

Panel A: Individual characteristics and types						
Type	Gender [1=male]	Wage employment indicator	Self-Employment indicator	Above Median Reservation Wage	Above Median Reservation Profit	At least a bachelor's degree
Type 1	.797	.665	.147	.265	.298	.759
Type 2	.845	.709	.201	.449	.426	.838
Type 3	.771	.676	.150	.336	.305	.740
Type 4	.777	.681	.144	.334	.373	.765
p-value	.051	.747	.224	<0.001	<0.001	.016

Panel B: Labour market experience and types					
Type	Employment	Permanent employment	Unemployed	Number of transitions	Management Position
	Years	*Years*	*Years*	*Count*	*Share*
Type 1	5.226	3.541	.577	.887	.109
Type 2	5.522	3.906	.329	.659	.192
Type 3	5.306	3.671	.456	.851	.124
Type 4	5.231	3.626	.510	.939	.139
p-value	.298	.094	.024	.070	.005

Notes This table describes the average characteristics of individuals (Panel A) and labour market experience (Panel B) of each type. Specifically, in Panel A this includes their gender (1 indicating male, 0 female), a dummy for their wage- and self-employment status, the probability they have an above-median reservation wage and profit based on data collected before respondents' attended the studio, and a dummy for whether or not they have included at least a bachelor's degree (BA, BSc, MA, MSc or PhD). Note that the median splits on reservation wage and profit do not yield a 50/50 split due to bunching in the underlying data at ETB10.000. Panel B includes the number of years they have been in employment including both self- and wage-employment, the number of years they have been in permanent employment, the number of years they have been unemployed, the number of labour market transitions they have gone through, and finally whether they were in a management position immediately before participating in the studio experiment. These numbers are calculated by assigning each individual to the pure type for which they have the highest estimated $\hat{\boldsymbol{\theta}}_i$. Then, a conditional average is taken for each pure type. To test whether intentions differ across types, we compute the Mahalanobis (Wald) statistic and compare it to a χ^2_3 distribution. This relies on the multivariate normal approximation to the posterior described in Section 4.1 of Gelman et al. (2013). For non-binary outcomes, we first create a binary split at the median to split the sample.

3.3.5 Labelling the types

For ease of reference, we now add labels to these four types. We stress that these labels reflect our suggested interpretation of the algorithmic output, and were not provided in advance to our unsupervised algorithm (as Ludwig and Mullainathan (2023, p.756) put it, 'The algorithm discovers, and people name that discovery'). In seeking these labels, we draw upon the classic work of Goleman (2000) – who described a set of archetypal leadership traits for large organisations. This has been a highly influential description of leadership for many corporate thinkers.[16] In his famous work, Goleman described six styles of *senior leadership*. We use this as a basis for describing four styles of *entry-level management* – the setting appropriate for our sample.[17] Specifically, we label the types, after estimation, as follows:

1. Type 1: The *authoritative manager* implements the rules set by the firm.
2. Type 2: The *affiliative manager* seeks shared ground.
3. Type 3: The *coercive manager* implements their personal objectives.
4. Type 4: The *timid manager* does not provide a clear explanation for their decision.

3.3.6 Robustness to attrition

About 60% of the original sample attended the studio. To compare the sample attending the studio with the original sample in the experiment, we run a number of descriptive logit estimations. In Appendix Table 3.24, we first examine whether some individuals were more or less likely to attend the studio. First, in column 1, we show that – compared

[16] This seminal piece had been cited almost 6000 times at the time of writing according to Google Scholar and featured in the Harvard Business Review's classic series of groundbreaking ideas in 2017.

[17] Goleman's six styles were: (i) coercive, (ii) authoritative, (iii) affiliative, (iv) democratic, (v) pacesetting and (vi) coaching. Of these, we use the coercive, authoritative and affiliative styles; to this, we add a 'timid' style that is evident in our results. Goleman's other three styles – namely, democratic, pacesetting and coaching – are primarily about *leadership* rather than *management*. In particular, Goleman describes those styles with their emphasis on coordinating teams and developing junior staff; these roles are not directly applicable to our management vignettes, nor to our sample of young professionals. Goleman's 'authoritative' leader was defined in terms of mobilising teams towards a vision; in the context of entry-level management, we repurpose this term to refer to enthusiasm to implement shared policy.

to the baseline sample – female participants are significantly less likely to attend, and that individuals with a BA degree are less likely to attend (although this is marginally significant). When we study attendance at the studio conditional on being in the four-year follow-up of the placement experiment (column 2), we similarly find that female participants are less likely to participate, and that individuals with an above-median reservation wage (conditional on employment) are less likely to attend. In Appendix Table 3.25, we use an Inverse Probability Weighting to show the robustness of our main Bayesian estimates to concerns about selection on these covariates.[18]

3.3.7 Managerial types and vignette content

Prior to recording each video, we asked each respondent how they intended to answer, and why. Table 3.2 describes the average prevalence of different answers to these questions, across the four identified types.[19] Specifically, we report prevalence of (i) distrust as a stated motivation for the planned response, (ii) relationship maintenance as a stated motivation for the planned response, (iii) intention to explain the decision in terms of fixed company procedure, (iv) intention to set an example for future cases.

These reported intentions align with our earlier interpretation of the four types. The affiliative type, for example, is least likely to distrust their counterparty, and much more likely to emphasise the value of maintaining an ongoing relationship; in doing so, they are far less likely than the other types to emphasise the importance of formal procedure. In contrast, the authoritative type stands out for its relatively high distrust of the counterparty and – above all – for being much more likely than the other three types to list formal procedure and setting of an example as key motivators. The coercive and

[18] We find no evidence for the hypothesis that individuals in the treatment group of our original placement experiment are more likely to attend the studio. In column (3), we study whether any individual characteristics are predictive of treatment status conditional on attending the studio. We find suggestive evidence that female participants who are treated appear to be less likely to decline the studio invitation; the sample appears to be well-balanced on all other characteristics. Importantly for our subsequent analysis, we find good balance on treatment status and whether either parent finished primary school.

[19] To generate summary statistics of types, we use the point estimates of $\hat{\boldsymbol{\theta}}_i$ to assign each individual to a model "pure type", to then assign them to a pure type T_i using: $T_i = \arg\max \hat{\boldsymbol{\theta}}_i$. Formal inference uses the full distribution of the posterior distribution of $\boldsymbol{\theta}$ to account for the issues with two-stage inference highlighted in Battaglia, Christensen, Hansen, and Sacher (2024).

timid types are similar to each other in their stated intentions, and both place a higher emphasis on formal procedure than the affiliative type. In Appendix 3.7.D, Table 3.10 provides a breakdown of the average prevalence of these intentions by vignette.

Table 3.2: Reported intentions by type

Type	Expresses distrust counterparty	Maintain relationship	Follow procedure	Set an example
Authoritative	.300	.259	.453	.332
Affiliative	.200	.332	.306	.250
Coercive	.274	.233	.361	.245
Timid	.257	.218	.347	.241
p-value	<0.001	<0.001	<0.001	<0.001

Notes: This table reports respondents' *ex-ante* intentions, elicited through open-ended questions before they answer each vignette. For each person we record whether they (i) say they do not trust the other party, wish to maintain a good relationship, mention following procedure, and aim to set an example for future interactions. After estimating our model we assign each respondent to the *modal* latent type – the topic k with the highest posterior weight $\hat{\theta}_{ik}$. The reported means are the sample means of the four intention indicators, conditional on latent type. We compute the Mahalanobis (Wald) statistic and compare it to a χ^2_3 distribution. This relies on the multivariate normal approximation to the posterior described in Section 4.1 of Gelman et al. (2013). The resulting p-value appears in the final row. Uncertainty from the first-stage LDA estimation is thus propagated via posterior draws.

3.4 Firms

3.4.1 Managers' preferences over management styles

Do these differences reflect alternative approaches to management that may be valued differently by different kinds of firms (as in Bandiera et al. (2020)), or is one type clearly preferred by firms? In this section, we turn to the results from our incentivised elicitation of firm preferences over the vignette responses of our respondents. We begin, descriptively, by noting that there is substantial idiosyncratic variation in firms' assessments of the same vignettes: in Appendix Table 3.29, we show that the probability that any two managers agree in their assessment of the same pair of candidates is about 60% (and that this figure is quite stable across vignettes, in both the wage employment and self-employment domains). In Appendix Table 3.30, we show that this probability is also largely invariant to similarities between the assessing firms.

Our Augmented Latent Dirichlet Allocation describes K types of studio responses. We now want to estimate firms' preferences over these styles. We elicited firms' preferences in the form of rankings over sets of three responses; we therefore estimate using a Plackett-Luce model (Luce, 1959; Plackett, 1975). This model (sometimes known as the 'rank-ordered logit') is specified as follows. For firm f, the latent utility of choosing each candidate i in vignette v is:

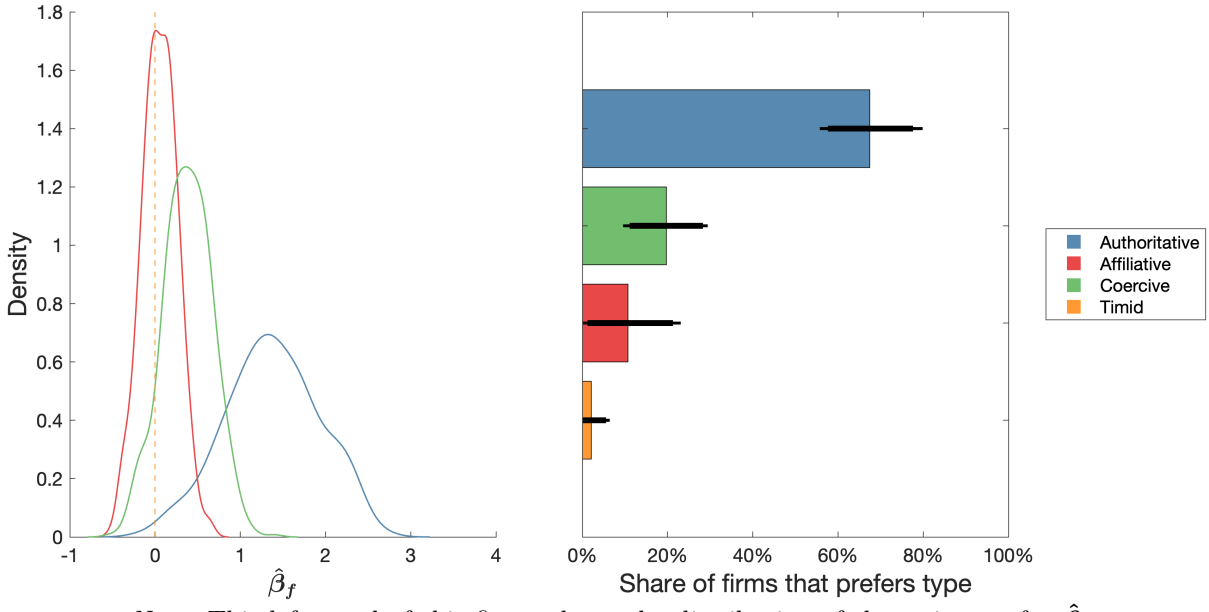
$$U_{fiv} = \boldsymbol{\beta}_f \boldsymbol{\theta}_i + \gamma_i + \varepsilon_{fiv}, \tag{3.7}$$

where θ_i is the vector of type probabilities from the Latent Dirichlet model,[20] and γ_i acts as a candidate-level random effect.We provide more detail on this model in Appendix 3.7.E.

We plot our estimates in Figure 3.4. The hierarchical model generates firm-level

[20] We jointly estimate the Latent Dirichlet model and the Plackett-Luce model. Joint estimation avoids the "attenuation" bias that would arise if we treated the type probabilities generated by the Latent Dirichlet model as observed data, which would result in both biased parameters and incorrect standard errors. Following Battaglia et al. (2024), this approach propagates the uncertainty in the Dirichlet parameters $\boldsymbol{\theta}_i$ and yields valid inference. See Appendix 3.7.E for details.

Figure 3.4: The distribution of the preferences for entry-level managers



Notes This left panel of this figure shows the distribution of the estimates for $\hat{\boldsymbol{\beta}}_f$ across all firms in terms of demand for entry-level managers. In this figure, the timid type is omitted and the estimates for $\hat{\beta}_{\text{Authoritative}}$, $\hat{\beta}_{\text{Affiliative}}$ and $\hat{\beta}_{\text{Coercive}}$ are plotted. The right panel shows that share of firms that most values a specific managerial type across our draws, i.e. the share of firms for which, for example, $\hat{\beta}_{\text{Authoritative}}$ is the largest element of $\hat{\boldsymbol{\beta}}_f$ for that draw. We calculate both the mean, and 90 and 95% credible sets based on our draws.

estimates for $\hat{\boldsymbol{\beta}}_{fs}$, and plot these estimates in the left panel (where the timid type is used as the omitted category); in the right panel, we show the estimated share of firms that first-preference each type. We find that the authoritative type is strongly preferred by most managers: indeed, about 70% of firms would first-preference that type. About 20% of firms instead first-preference the coercive type. Hiring managers have broadly similar preferences for the timid and affiliative types, and very few firms would first-preference either type.

To interpret these results, we consider what would happen if we shift the value of $\boldsymbol{\theta}$ to represent a 'pure' type manager for one member of each triplet. Specifically, for the simulation of the parameter $\beta_{\text{authoritative}}$, we adjust one individual in each triplet to have $\theta_{\text{authoritative},i} = 0.85$, setting the other elements of the vector $\boldsymbol{\theta}_i = [\theta_{\text{affiliative},i}, \theta_{\text{coercive},i}, \theta_{\text{timid},i}]$ to 0.05.[21] We then simulate the utilities of the HR managers using this updated $\boldsymbol{\theta}$ and calculate the new winning probability for this simulated individual. Initially, the win-

[21] This represents approximately the 95th percentile in the estimated distribution of $\hat{\boldsymbol{\theta}}$.

ning probability of each individual in the pairwise comparisons is 50%. This probability increases to 62% for the authoritative type, decreases to 45% for the affiliative and timid types, and slightly decreases to just under 50% for the coercive type. Another way of thinking of this is through a variance decomposition – which shows that there is substantial idiosyncratic noise, but that over half of the non-idiosyncratic variation in preferences can be explained by the preferences over our four types.[22]

In Appendix Figure 3.16, we examine the distribution of the parameter estimates for β_f for subsets of the firms. We split the sample (i) by sector (manufacturing versus services), (ii) by whether the manager has an above- or below-median score on the World Management Survey, and (iii) by whether the firm is public or private. We find no meaningful relationship between the distribution of β_f and any of these different dimensions.

We then estimate three extensions of the model, each offering greater flexibility than the main specification. First, in Appendix Table 3.26, we allow managers' preferences over types to vary at the vignette level (that is, we estimate β_{fv} rather than β_f). We find broadly similar preferences across all vignettes – in particular, a strong preference for the authoritative type in each of the vignettes – with some interesting heterogeneity in preferences for the coercive type (firms are more receptive to the coercive type in the scenarios where the firm clearly holds a dominant role in the relationship – namely, in vignettes featuring employee absence, a request for a pay rise, and the request from a client). Second, in Appendix Table 3.27, we allow managers' preferences over types to vary based on the gender of the young professional. We find no significant difference in preferences by candidate gender (the credible sets on the interaction include zero in all cases); looking solely at point estimates, we note that, if anything, the authoritative management style is particularly valued among female candidates.

[22] Specifically, we decompose the variance of the parameter estimates of the model specified in Equation 3.7. We calculate: (a) the variance of the total non-random component of the utility function, (b) the variance of just the $\boldsymbol{\beta_f \theta_i}$ term and (c) the variance of γ_i. We obtain $\widehat{\text{var}(\boldsymbol{\beta_f \theta_i} + \gamma_i)} = 0.63$, $\widehat{\text{var}(\boldsymbol{\beta_f \theta_i})} = 0.34$ and $\widehat{\text{var}(\gamma_i)} = 0.29$.

3.4.2 Robustness: Random variation in actors' gender

We previously described the five vignettes that were played to young professionals. As part of this experimental design, we additionally randomized both (i) the order in which young professionals viewed the five vignettes and (ii) the gender of the actor shown in each vignette (having recorded each vignette twice: once with a male actor and once with a female actor). In this final part of the analysis, we exploit this additional exogenous variation to generate a robustness check on our main results. Specifically, we exploit the random variation in whether the first actor viewed by a respondent is female.[23]

To analyse this variation, we run three kinds of analysis – all reported in Table 3.23. First, we test the effect upon managerial types of having seen a female actor in the first vignette. To do this, we use the Hamiltonian Monte Carlo chains, and compare the posterior distributions between those respondents who initially saw a male actor and those who initially saw a female actor; we then use these distributions to construct point estimates of the causal effect, and 95% Bayesian credible intervals. Panel A shows that exposure to a female actor in the first vignette leads respondents to rely more on the authoritative management style and less on the coercive style (with a 95% credible interval that excludes zero for both effects).

In Panel B, we report a series of OLS regressions to test the average treatment effect on measured attributes of professionals' responses; we estimate:

$$Y_{ive} = \alpha + \beta \cdot \texttt{First\_Actor\_Female}_i + \gamma_v + \lambda_e + \varepsilon_{ive}, \tag{3.8}$$

where we include vignette (γ_v) and enumerator (λ_e) fixed effects, and cluster errors at the individual (i) level. We examine a set of four outcome variables: a binary indicator for relying on the firm's interest as justification; an indicator relying on the respondent's own interest as justification, an indicator for relying on a formal policy as source of authority,

[23] We can conduct a similar analysis using the gender of each separate actor – but this analysis must then proceed at the level of the respondent-vignette pair. Given our particular focus on managerial types – that is, a property that we estimate at the level of the respondent – it makes more sense to use this simpler respondent-level variation.

and an indicator for relaying on seniority as source of authority.[24] We restrict the sample to data from the second through fifth vignettes a respondent sees. We find that, where the first actor is a woman, respondents are more likely to rely on the firm's interest instead of their own interest as their justification, and more likely to rely on formal authority instead of seniority. This confirms the result in Panel A: that, on average, respondents exhibit more authoritative behaviour after being exposed to a female actor in the first vignette.

Finally, we test the impact upon firms' responses – stressing that the HR managers were never told whether the young professionals had seen a male or female actor. To assess this, we estimate a Plackett-Luce model with a single covariate; we define the latent utility of firm f's assessment of candidate i in vignette v as:

$$y^*_{fiv} = \beta \cdot \texttt{First\_Actor\_Female}_i + \varepsilon_{fiv} \tag{3.9}$$

Panel C of Table 3.23 reports the results from this exercise. Specifically, we calculate the implied probability that an HR manager ranks a candidate who first saw a female actor more highly than one who first saw a male actor. We estimate that seeing a female actor first increases by about 5.2 percentage points the probability that an HR manager would prefer the candidate as a prospective entry-level manager, and by about 6 percentage points the probability that an HR manager would prefer the candidate as a prospective entrepreneur; the credible sets for each of these estimates excludes zero.

In sum, the randomisation of actors' gender generates additional excludable variation in respondents' measured managerial traits – and we show that, consistently with our main results, firms prefer managers exhibiting the 'authoritative' type.

[24] These specific variables are selected to illustrate the shift from a coercive to an authoritative management style, we report a series of multinomial logit regressions of the effect of seeing a female actor on the full set of attributes from which these were selected in Appendix 3.7.L.

Table 3.3: Actors' gender and management styles

	Panel A: Effect on estimated types			
	Authoritative	Affiliative	Coercive	Timid
First actor female	0.030***	-0.004	-0.030***	0.003
Constant	0.262	0.191	0.296	0.251
Bayesian Credible Interval	[.012 .049]	[-.021 .013]	[-.051 -.006]	[-.021 .031]
N	982	982	982	982

	Panel B: Effect on selected attributes (Full set in Appendix 3.7.L)			
	Justification		Authority	
	Rely on firm's interest	Rely on own interest	Rely on formal authority	Rely on seniority
First actor female	0.032*	-0.028**	0.021	-0.026
	(0.019)	(0.014)	(0.015)	(0.020)
Constant	0.554***	0.178***	0.154***	0.537***
	(0.013)	(0.010)	(0.010)	(0.014)
Enumerator FE	Yes	Yes	Yes	Yes
Vignette FE	Yes	Yes	Yes	Yes
Mean dep. var	0.552	0.168	0.163	0.511
N	6887	6887	6887	6887

	Panel C: Effect on managers' assessments			
	Ranking Data		Normalised likert score	
	Manager	Entrepreneur	Manager	Entrepreneur
First actor female	0.052**	0.060***	0.083**	0.093***
	[0.010 0.093]	[0.017 0.102]	(0.033)	(0.033)
Constant	0.474***	0.470***	-0.045*	-0.050**
	[0.453 0.493]	[0.449 0.492]	(0.025)	(0.025)
Vignette FE			Yes	Yes
N			6874	6869

Notes: This figure displays the causal relationship between the first actor a respondent sees and their subsequent responses (for the second to fifth vignette). Panel A shows a causal effect on adopting a more authoritative management style (with 95% Bayesian credible intervals in square brackets). In Panel B we report the effect of seeing a female actor on selected attributes exhibited by the respondent in subsequent vignettes based on a linear regression. These attributes are selected to illustrate the shift from a coercive to an authoritative management style; Appendix 3.7.L reports a sequence of multinomial logit regressions on the full set of attributes from which these were selected. The third panel shows the effect of the first actor being female on the HR managers' rankings. The constant is the probability of winning each pairwise comparison after seeing a male actor in the first vignette, with the 95% credible set reported. The parameter "First actor female" is the coefficient β in probability space, i.e. the mean marginal effect, again with the 95% credible set in square brackets. The latter two columns use the non-incentivised, normalised Likert score that the HR managers give each candidate. We implement a simple OLS regression with vignette fixed effects with standard errors clustered at the HR manager level. Statistical significance (in the classical sense for regressions, and testing whether the 90, 95 or 99% credible set contain zero for Bayesian analysis), is denoted where appropriate by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

3.4.3 Workers' preferences for management styles: A model-informed field survey

The previous subsections show a clear preference from firm managers for the authoritative trait. To what extent is this preference shared by firm employees? On the one hand, for example, employees might share managers' preference for the authoritative trait – perhaps because they respect the way that the authoritative type communicates clearly on the basis of set rules. On the other hand, employees might prefer – for example – the more amiable interpersonal style of the affiliative type.

To answer this question, we returned to the field for a series of semi-structured interviews with firm employees. We did this after obtaining our model estimates – so that the interviews could focus on interpreting our model results. Specifically, we interviewed a total of 66 employees, drawn from 17 firms in our original sample; we did this through a series of 12 focus groups in Addis Ababa, in May 2025. The average worker was 31 years old and had spent four years with their firm, which were mostly operating in manufacturing (51.5 percent) and services (42.4 percent). Each respondent viewed two vignettes: the vignette concerning line-management of an employee, and the vignette concerning a requested pay rise. (We chose these two vignettes for their direct relationship to employee management.) For each of these two vignettes the respondent then watched the responses of five different young professionals (each of whom had given permission for their video to be viewed in this way). These five young professionals were chosen based on the model estimates; specifically, we chose (i) a professional close to being a pure authoritative type, (ii) a professional close to being a pure affiliative type, (iii) a professional close to being a pure coercive type, (iv) a professional close to being a pure timid type, and (v) a professional chosen for being approximately an equal weighting of those four types. (That is, in the context of the tetrahedron illustrated in Figure 3.3, we chose young professionals from near to each of the four vertices and from close to the centroid; we illustrate this in Appendix Figure 3.17.) We used a total of 60 young professionals (that is, 12 in each of these five categories), to ensure that our interview

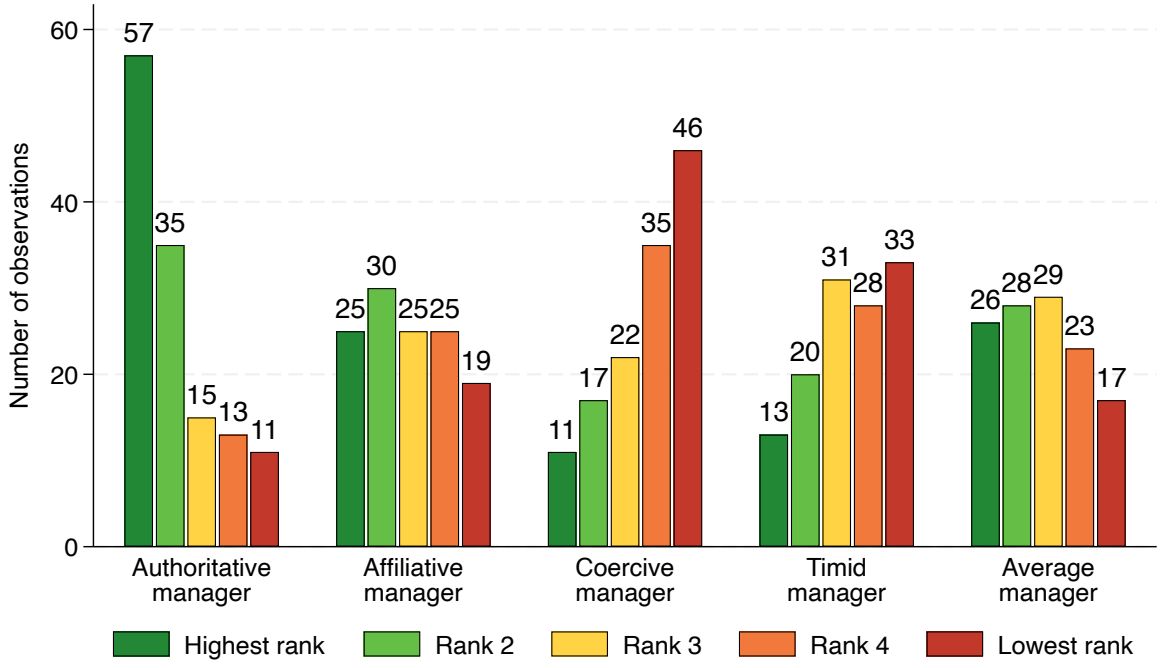
results are not driven by the idiosyncrasies of specific individual videos. Further, we ensured that no employee saw the same professional more than once, and that each employee either saw only male professionals or only female professionals. We used neutral labels to refer to all of the videos.

We then asked each employee a series of questions about the suitability of each young professional for their firm. Specifically, we asked each employee to provide a Likert-scale assessment of (i) whether the young professional would enhance firm productivity, (ii) whether the young professional would help the firm to achieve its targets, (iii) whether the employee would be happy to work hard for the young professional, (iv) whether the young professional would treat different employees equally, and (v) whether the style of response is typical for the employee's firm. Finally, for both vignettes, we asked each employee to rank the five professionals they had seen in order of their likely contribution to the long-term success of the firm.

Figure 3.5 shows the distribution of these overall rankings. The figure shows a clear and significant preference for professionals of the authoritative trait. In Appendix Table 3.33, we find the same clear pattern across all of the questions asked: the authoritative professionals were significantly preferred to all other professionals across each of the questions asked. Qualitative follow-up responses placed particular weight on employees' perceptions that they would be treated fairly and consistently; Appendix Table 3.33 shows a clear preference for the Authoritative type in this regard, and the qualitative discussion suggests that this was a central reason for the strong overall assessment of authoritative professionals. Indeed, in this regard, it is interesting to note that the affiliative professionals were the least preferred type across all questions asked – including in employees' perceptions that such professionals would treat employees equally. Qualitative follow-up discussion suggested that – far from seeing this type as being their ally or supporter – employees may have viewed affiliative professionals as being overly impressionable and prone to favouritism. In Appendix Tables 3.34 and 3.35, we disaggregate by vignette; we find that these patterns are very similar across the two vignettes (with the timid professionals performing relatively better in the line management scenario than in the pay rise

scenario).

Figure 3.5: Distribution of ranks by management style



Notes: This figure summarizes 132 rankings of 66 workers of the effectiveness of our young professionals as managers. For each "pure" type, we count how many times that type is ranked highest, second, third, fourth, or fifth and display these counts here. We conduct all statistical inference for this section in Appendix Table 3.33 using a rank-ordered logit model. We note that the authoritative type is significantly preferred over all other types ($p \leq 0.001$ in a binary comparison of the coefficients). Next, the coercive and average type are similarly preferred over the affiliative, and timid types, and finally the timid type is preferred over the affiliative type.

3.5 Experimental variation in managerial exposure

3.5.1 The causal effects of managerial experience

The results in section 3.4 show that authoritative managers are preferred by firms; the earlier results in section 3.3 show descriptively that authoritative managers also have historically greater labour market attachment and higher parental education. Together, these results suggest that managerial traits may be a mechanism for persistence in labour market exclusion – but do not tell us whether labour market experience influences the development of managerial traits, or whether managerial traits, instead, determine labour market outcomes. To answer this question, we turn to evidence from respondents' partic-

ipation in a management experience experiment five years prior to attending the studio.

As noted earlier, our sample was drawn from a set of young professionals who previously participated in an experiment in which a random subset was assigned to a management placement. Specifically, that earlier experiment (conducted in 2016 and 2017) offered young professionals a one-month placement shadowing middle managers in their daily work in medium and large firms; this experiment is discussed in detail in Abebe et al. (2024). This sampling strategy allows us to estimate the causal effect of labour market exposure on different management styles among young professionals. For each management style – timid, authoritative, affiliative, and coercive – we compute the treatment effect as the difference between the estimated latent trait for treated versus untreated individuals. We calculate the average treatment effect as:

$$\Delta\theta_i^{(d)} = \theta_{i,\text{treated}}^{(d)} - \theta_{i,\text{control}}^{(d)}, \tag{3.10}$$

where $\Delta\theta_i^{(d)}$ represents the treatment effect for style i for Hamiltonian Monte Carlo draw d. For each category, we report the mean and the 95% credible interval of the distribution of these draws. We separately conduct this exercise for three groups: the full sample, individuals whose parents did not complete primary school, and those with at least one parent who did.

We find, in Table 3.22, that the management experience treatment makes individuals on average more authoritative and less timid (column 2), and that this result is entirely driven by individuals whose parents have not finished primary school (column 4). We find no evidence for any treatment effect for individuals whose parents finished primary school (column 6).

We draw two conclusions from this result. First, management experience shapes an individual's management style. Treated individuals became significantly more likely to display an authoritative style of management, and less likely to display a timid style. Since the treatment focused on shadowing a manager, this suggests that the intervention shows what effective management looks like rather working through "learning by doing". Second, we find that this average treatment effect is *entirely* driven by participants for

whom neither parent completed primary school. We interpret this finding as showing that these individuals do not learn the established norms from their parents, as shown by the lower weight on the authoritative management style in the absence of the treatment, but are able to learn these norms through exposure to formal management practices at an established firm. Strikingly, the treatment significantly reduces the average difference in management styles between individuals for whom at least one parent completed primary school and for those for whom neither did so - from 7.7 percentage points for untreated individuals to 2.0 percentage points for treated individuals. These findings suggest that managerial experience plays an important role in shaping managerial traits, particularly for individuals with less prior exposure to formal firms.

Table 3.4: The causal effect of managerial experience on management style by parents education

	Full sample		Low parental education		High parental education	
	(1)	(2)	(3)	(4)	(5)	(6)
Authoritative (%)	26.5	2.4 [0.2, 4.5]	22.5	5.4 [2.4, 8.7]	30.2	-0.3 [-3.4, 2.6]
Affiliative (%)	18.4	1.2 [-0.5, 2.9]	19.1	1.0 [-1.6, 3.6]	17.7	1.4 [-1.1, 4.0]
Coercive (%)	28.0	-0.1 [-2.3, 2.3]	29.0	-1.2 [-5.0, 2.1]	27.1	0.9 [-2.5, 4.2]
Timid (%)	27.1	-3.6 [-5.8, -1.5]	29.4	-5.3 [-8.3, -1.8]	24.9	-2.0 [-5.3, 1.1]
N	479	500	229	239	250	261

Notes This table reports the treatment effect of the management experience experiment on the managerial traits of individuals. The treatment effect is calculated based on the distribution of the difference in the average value of $\boldsymbol{\theta}$ for treated and untreated individuals. Columns (1), (3) and (5) report the average estimated value of $\boldsymbol{\theta}_i$ for individuals that were not treated in the management experience experiment for respectively all individuals, individuals whose parents did not finish primary school and for individuals for whom at least one parent did. Columns (2), (4) and (6) report the treatment effect of the management experience experiment on their managerial traits for these three groups respectively. In columns (2), (4) and (6) both the average treatment effect and the 95% credible interval, in square brackets, are reported.

3.5.2 The causal effect on managers' assessments

We next test whether this experimental managerial exposure affects how candidates are ranked by HR managers. We estimate a Plackett–Luce (rank-ordered logit) model on firms' ranking data and translate posterior draws into pairwise winning probabilities. For firm f, the latent utility of choosing each candidate i in vignette v is:

$$U_{fiv} = \delta_1 \cdot T_i + \delta_2 \cdot \texttt{PE}_i + \delta_3 \cdot T_i \cdot \texttt{PE}_i + \varepsilon_{fiv}, \tag{3.11}$$

where T_i is a dummy equal to one if the individual is treated, PE is a dummy equal to one if either parent has completed primary school and ε_{fiv} is a Type-1 EV distributed error term. We estimate this model in Stan.

To interpret these results, we use the posterior draws to estimate winning probabilities for various pairwise group comparisons, keeping fixed either parental education or treatment status. In Table 3.5 we first show that, across the full sample, a treated individual has a 53.3% chance of winning in a binary comparison with an untreated individual – suggesting the effect of the treatment on managerial styles is reflected in HR managers' assessments. Next, we examine the treatment effect conditional on parental education. For individuals whose parents did not complete primary school, a treated individual has a 55.1% chance of winning when compared to an untreated individual; for individuals whose parents did complete primary school this is equal to 51.3%, which is not estimated to be significantly different than 50%. Finally, we show that, in the absence of treatment, individuals with low parental education are at a disadvantage; however, conditional on treatment, parental education no longer affects managers' assessments.

Table 3.5: The causal effect of managerial experience on managers' assessments

Comparison	**Winning Probability (%)**
Overall Treatment Effect	
Treated vs. Untreated (All)	53.3***
	[51.6, 55.0]
By Parental Education	
Treated vs. Untreated (Low parental education)	55.1***
	[52.6, 57.4]
Treated vs. Untreated (High parental education)	51.3
	[49.0, 53.7]
Gap in Absence of Treatment	
Low vs. High Parental Education (Untreated)	45.3***
	[42.8, 47.7]
Gap Conditional On Being Treated	
Low vs. High Parental Education (Treated)	49.7
	[46.3, 51.0]

Notes: Entries report posterior means of pairwise "winning" probabilities from a Plackett–Luce (rank-ordered logit) model estimated in `Stan`. For any two profiles i and j, the winning probability is $\Pr(i \succ j) = \frac{\exp(U_{fiv})}{\exp(U_{fiv})+\exp(U_{fjv})}$, computed at each posterior draw and then averaged. Brackets show the 95% credible intervals calculated as the range between the 2.5th and 97.5th percentile of the distribution of the posterior draws. We show the heterogeneity of this treatment effect by parental education, and the remaining gap by parental education conditional on treatment status. *Treated vs. Untreated (All)* reports the treatment effect for the full sample. *By Parental Education* conditions on low vs. high parental education (low means neither parent completed primary school; high mean at least one parent completed primary school). *Low vs. High Parental Education (Untreated/Treated)* compares otherwise identical profiles that differ only in parental education, holding treatment status fixed. A value of 50% indicates no difference in expected rank in a binary comparison.

3.6 Discussion

In this paper, we have developed a novel experimental approach to assessing managerial traits – using a controlled studio setting to observe and analyse how young Ethiopian professionals respond to realistic workplace challenges. By combining experimental data with Bayesian hierarchical modeling, we identified four distinct managerial types, which we label as 'authoritative', 'affiliative', 'coercive', and 'timid'. This exercise has generated several key insights.

First, we show meaningful heterogeneity in managerial traits: our four archetypes reflect meaningful variation in how young professionals approach decision-making and conflict resolution in managerial roles. The authoritative type, for instance, emerges as the most dominant managerial style, particularly valued by firms for entry-level managerial roles. This type tends to emphasise formal procedures, personal authority, and firm-centered interests – contrasting sharply, for example, with the affiliative type – who seeks shared ground and often concedes to a counterparty.

Second, our analysis reveals that firms in Ethiopia, across a range of industries, tend to prefer authoritative-type managers for entry-level positions. Some previous studies (e.g. Bandiera et al., 2020) have emphasized the heterogeneity in firm preferences for senior managers; our results suggest that firms may place greater value on uniformity in managerial traits at the lower levels of management. This homogeneity could reflect the need for clear, consistent, and rule-based decision-making among less-experienced managers. This result accords with several anecdotal characterisations of large Ethiopian firms – which tend to emphasise the importance of strong centralised leadership, with a clear role for corporate hierarchy.

Third, we show that young professionals with more exposure to the labor market are substantially more likely to exhibit traits of the authoritative archetype. Our evidence based on random assignment to a previous management placement experiment suggests this is at least partially a causal link. This finding suggests that managerial styles are not purely innate, but can be shaped and developed through professional experience.

These results align with previous studies showing the importance of early exposure to managerial environments in shaping long-term behavioural traits (Bertrand and Schoar, 2003; Malmendier et al., 2011).

Methodologically, our paper demonstrates the feasibility and the value of using studio-based vignettes to assess managerial traits. Unlike traditional observational studies, our controlled setting allows us to systematically compare managerial responses across a fixed set of scenarios, removing the endogeneity inherent in real-world managerial decision-making. This approach opens up new avenues for studying managerial behavior – particularly in contexts where direct observation of managerial actions may be impractical or biased by the organizational setting.

More generally, our study complements the growing literature on labor market exclusion faced by young professionals. The finding that authoritative-type managers are both in higher demand by firms and more likely to have greater labor market exposure suggests that managerial traits may serve as a key mechanism by which some young professionals achieve sustained labor market success, while others face exclusion. Our results on the relevance of parental education speak particularly to this issue. Respondents with less educated parents are significantly less likely to exhibit authoritative-type traits than respondents with more educated parents, and are significantly less likely to be preferred by firms (notwithstanding that firms have no direct means of observing respondents' parental education). However, the exogenous variation in labour market exposure essentially closes completely both of these gaps.

This suggests several novel angles for policy innovation. Policy to support disadvantaged jobseekers has traditionally focused on development of human capital, or on effective signalling of existing skills (Abebe, Caria, Fafchamps, Falco, Franklin, and Quinn, 2021). Our paper suggests a third possibility: working with young professionals to develop the kind of authoritative professional traits sought by large firms. Our results show that spending time in established firms is one mechanism to achieve this; there may

be several others.[25]

Future research could expand on this work by exploring how these preferences evolve as managers move into higher-level roles and by investigating the role of cultural and institutional factors in shaping managerial behavior. Additionally, the scalability of our experimental method presents opportunities for further research across different contexts and labor markets, providing a valuable tool for understanding the development of managerial talent in a wide range of settings.

[25] In a related space, several initiatives seek to support disadvantaged jobseekers by targeting interview appearance, among other skills. For example, 'Dress for Success Worldwide' provides 'professional attire services' as part of its support for women jobseekers; in Ethiopia, 'The Talent Firm' is a professional recruitment company that provides 'The Professional's Closet': a service providing donated professional attire for professional jobseekers.

3.7 Appendices

3.7.A Vignette scripts

3.7.A.1 Respondent management vignettes

Vignette 1: An absent employee

NARRATOR: *In this video, an accounting clerk who has been absent for three days without advance notice to her immediate boss is shown. (S)he shows up for work on the fourth day, explaining that (s)he was unwell. Do watch the video carefully.*

THE CLERK: I realize that I did not come to work for the last three days. I could not also give you any advance notice as this was something beyond my control. I was not feeling well. I had a bad flu that made me see a doctor. The doctor recommended that I should take it easy and perhaps take a few days of rest at home. I thought that, with my conditions, coming to the office and working on our accounts would be difficult [actor coughs slightly]. Also, I did not want others to catch my flu as well, so I stayed away.

Boss, you remember Almaz, my colleague, who had a bad flu three weeks ago. Instead of staying at home, she chose to come to work while being very sick. The next day, almost half of the staff in our department had a terrible headache and were coughing. They caught whatever she had. I certainly did not want that to happen [actor shakes head in the negative].

Vignette 2: Negotiating with a supplier

NARRATOR: *Here we have a supplier, who explains to you that (s)he cannot fulfil an order according to specifications, because of problems sourcing input materials. (S)he offers to supply a replacement of inferior quality instead. Note that*

quality is the most important feature of the firm's product and that the firm is known in the market for producing and selling the highest quality products in the industry. Do watch the video carefully.

THE SUPPLIER: As you know, the main roads that we use to transport our shipments from Djibouti have been blockaded in the past several weeks. It has been three weeks since we got our last shipment of the materials that we promised to supply to you. During the course of the past three weeks, our inventory of these materials has been depleted. I am thus sorry to inform you that we cannot supply you with the materials as specified in the original supply agreement.

However, the good news is that we have a pretty good stock of alternative materials that we could supply you at a cheaper rate. We cannot, however, guarantee that these materials currently in stock meet the quality and specification requirements that were stipulated in the original supply agreement. Nevertheless, I think what we have for you is a good deal that is offered with the cheapest possible price. Of course, once our sourcing problem is resolved, we will continue to supply you with materials that comply with the specifications that you set.

Vignette 3: An employee seeks a pay rise

NARRATOR: *In this video, we will show you a production worker, coming to you and asking for a pay rise. Your company does not have a plan to increase any of your workers' salary this year. Do watch the video carefully.*

THE PRODUCTION WORKER: Boss, you know how dedicated I am to my work. In the past three months, our department has done very well and we have met the quarterly targets set by the management. I'm very happy about that.

As you know, I have been the most productive worker in the department. I am one of the few people in the department that exceeded the personal productivity targets that you set for us at the beginning of the year. Of course, I also did not have even a single day of absence recorded in the completed quarter. If you don't believe me, you can check the attendance repository and the productivity (or KPI) sheet. As you know, I also worked for the company for the past ten years, not so many people can say that! I am not sure whether you know my mother, Feleku Kuma; she is widely known around here, she also worked here for a very long time.

I think these speak volumes to my dedication and loyalty to this company. But honestly, I don't think my current pay reflects the effort that I exert and the loyalty that I and my family have for the company. In fact, I spoke to people who work with our competitors at the same type of occupation as I am; surprisingly, their pay is much larger than mine. I am sure you agree with me that I deserve a pay rise. Can you let me know as to when and how much I should expect to get as a pay rise?

Vignette 4: Negotiating an adjustment with the bank

NARRATOR: *This video shows a bank manager, who calls to remind you about an unpaid loan installment. But you will not have sufficient funds to pay the bank for another two weeks. Take a watch.*

THE BANK MANAGER: Thanks a lot for taking my call. As you know, you are one of our clients who missed the latest installment on your outstanding loan from our bank. Our bank treats its customers with utmost respect and encourages timely payment to avoid breaches to loan terms and conditions. But also, we have quite a lot of pressure from the board of directors to ensure that scheduled loan payments are made on time. Failure to pay loan installments not

only increases the interest payment you owe us; it also reduces you credit rating with us. Next time you ask for loan, you may be denied access altogether – or may be eligible to only a fraction of the amount that you have applied for. I thus cannot stress enough the importance of settling your outstanding installment as soon as you can. That said, I am wondering whether you will be able to pay the unpaid loan instalment very soon. I will follow-up this call with a written reminder to alert you to the periodic payment that we expect to be paid from your company.

Vignette 5: Negotiating with a client

NARRATOR: *The next video shows a client of your business, who has not paid what you have invoiced, in spite of one reminder. The client comes to place a new order.*

THE CLIENT: I really like the service I got from your company last time. I don't think I had the chance to thank you for that – so thank you. Oh, I did not forget about the payment and I confirm that I have received the invoice you issued. But you know how it is these days, business is not great, and our product is not exactly flying off the shelves. We are really short on cash at the moment. I will thus have to ask you to bear with me a little to pay for the previous order. With the risk of stretching your generosity, I'm here today to place another order. In a spirit of our long-standing relationships, I appreciate if you could deliver my new order soon without waiting for payment for the previous order. I promise we will meet our payment obligations for the previous and new orders once we sort out our cash flow problem.

3.7.A.2 Assessment of management styles: Hypothetical workplace conflict

Setting up the scenario

NARRATOR: You are faced with a situation where workers go on strike demanding proper latrines. A representative of the workers comes to talk to you. Take a look at the video carefully.

THE WORKER REPRESENTATIVE: In my capacity as a representative of workers in this factory, I would like to speak to you about the ongoing strike in the factory. I am sure that by now you realize what workers are striking about. But with the slow response from the line supervisors and the deafening silence from you, I am not certain that the management in the factory understood the severity of the problem. That is why I'm here.

Look, the strike is about the poor toilet facilities in the factory. As you recall, during the worker/management meeting, many workers complained about the lack of sufficient toilet facilities; there are very few toilets in the compound and workers have to walk to the other end of the compound to access them. Even when they reach the toilets, they have to wait their turn, sometime for several minutes, to use them. In recent months, the problem has gotten worse. The few toilets that are available either do not have water or the flushing system does not work and tissue papers are not available. To make things worse, sometimes the toilets are locked and workers are not told why they are locked and who is keeping the keys. Further, even when workers try to use the toilets, they face restrictions on the number of toilet breaks they are entitled to take.

The attitude of your supervisors towards toilet use is also appalling. Our female workers suffer a great deal particularly when they are experiencing their period. Studies indicate that failing to use toilets on the course of a day lead to several health problems including damage to the bladder and several types of infections that have serious implications on the workers' wellbeing. We as

workers are very much aware of the implications of poor toilet facilities. Please do not underestimate our intelligence, we know about all this and our rights. Through our discussion with fellow workers, we came to the conclusion that our voice needs to be heard. We are particularly troubled by the dominant view of the management that workers shirk using toilet breaks as an excuse. Such views are not only unfounded, we reject the thought process behind it as it deprives the dignity of workers as human beings. Unless our demand for improved toilet facilities is not met, our strike will continue.

Six possible responses to the scenario

We then use different actors to record six separate possible managerial responses. The responses are designed to represent different styles of engagement. In this summary, we suggest several descriptors for each managerial response; these descriptors were not provided to the respondents.

NARRATOR: After listening to the complaint by the worker representative, six different managers respond in different ways. Please do watch the videos.[26]

MANAGER 1 [*APOLOGETIC; COLLABORATIVE; EMPOWERING*]: We understand workers' complaints about the poor state of toilet facilities in our factory. We admit that we have been slow to act and hence invited the strikes onto ourselves. On behalf of the company, I would like to apologize for our failure to nip this problem in its bud before it reached to this level. We will take immediate actions to rectify the situation. For a start, we will constitute a committee that will include workers' and management representatives. The committee will work on identifying the key problems in the existing toilet facilities and would suggest solutions for improving access and quality of toilet facilities. In

[26] For clarity, these are ordered based on the senior managers' ranking which was used to incentivise the respondents. Respondents' instead saw these in a random order.

parallel, the management will meet and earmark the necessary resources required to improve the facilities. We actually want the whole process to be led by workers and their representatives. We believe that there is no substitute for better working conditions to get the best out of our workers. With this pledge, we would like you to convince the workers to go back to work.

MANAGER 2 [*ACKNOWLEDGING; PROACTIVE; REALISTIC*]: Honestly, this is a fair request. We should have seen this coming. As we promised during the worker/management meeting, the management takes all workers' requests very seriously. We are now making preparations to invest in improving the toilet facilities. Some of the improvements related with maintaining the flushing system of the existing facilities can be readily made. But increasing the number of toilets and digging septic tanks could take longer time. As you know, obtaining construction permit in this part of the city takes quite a long time. So even if we want, we may not be able to overhaul the whole facility at once. We thus ask for patience on the workers end. With this in mind and with the promise to make feasible changes immediately, we request you to stop the workers' strike at once.

MANAGER 3 [*APPRECIATIVE; CONCILIATORY; BLAME-SHIFTING; BEHAVIOURAL FOCUS*]: Thanks for bringing this to our attention before things go out of hand. We really appreciate your effort to be a voice of workers. The way your represent the concerns of your workers has been exemplary – and, on behalf of the company, I would like to express my gratitude. Of course, even if you are the representative of workers, we always felt that you are one of us, the management team. As both a worker representative and as someone who works closely with us, you understand how it has been difficult to manage workers, particularly their use of toilet facilities. We agree that facilities should improve, and that can be done.

But this issue is not that important. What worries us the most is workers have been abusing their toilet breaks by sitting idle and chattering with fellow workers during the breaks. Some of them spend more than 50 minutes inside the toilet, locked inside with their phone, checking Facebook or playing games. We told them repeatedly that this is not an acceptable behavior and it ought to change. When they refused, we started locking some of the toilet rooms. I think you should help us to get their act together. We can work together to change their behavior use their rightful toilet breaks only for intended purposes. Explain to them that the strike should come to an end and they should stop abusing their breaks. We can then discuss about ways to improve the quality of the facilities.

MANAGER 4 [*ACKNOWLEDGING; DEFLECTIVE; POSTPONING; SUBTLE PRESSURE*]:

We understand that toilet facilities are important and workers have the right to complain about problems in accessing these facilities in our compound. But everyone knows that we are struggling at the moment: business is not good and we are making losses. Asking the management to invest in improving the toilet facilities at the moment is not a good idea. This simply is not the right time to do that.

As a representative of workers, you should be aware that management does not like to be ambushed or cornered into making decisions in difficult times. I am sure you also do not want to take the blame for creating trouble in the company while conditions are already hard on us. I know you are also concerned about your career prospects. This is really a tough time for everyone, so let us weather the storm together and we can resume the discussion about improving the toilet facilities afterwards.

MANAGER 5 [*DISMISSIVE; THREATENING; DEFENSIVE*]: As you said, we have discussed this during the worker-management meeting. During the meeting, we were very clear. We can only meet demands that are convincing and consistent with the company's policy. We told the workers that we know what kind of toilet facilities other employers provide for their workers.

Did you have a chance to visit other company's toilet facilities? If you or your colleagues had done that, you would have realized that ours is not of sub-standard quality compared to them. If anything at all, we provide flushing toilets and fresh water, which many don't. We also know the kind of toilet facilities workers have at their place of residence, mostly pit latrines that are unsafe and often dirty. We certainly provide better facilities than those at home. So if you don't like our toilets here, you and your comrades can go and work elsewhere. And believe me, this is what is going to happen if you don't cease striking in the next 24 hours.

MANAGER 6 [*LEGALISTIC; HOSTILE; THREATENING; DISMISSIVE*]: Who do you think you are, barging into my office and putting unwarranted demands on this company? I am not going to discuss about toilets or about any related issues with you. No, not at all. Forget toilets, let us talk about the strikes. Do you know what the labor law says about strikes? I am sure you don't! Had you known, you wouldn't be striking! Let me enlighten you. You see this [actor shows a copy of a labour proclamation]: this is Labour Proclamation No. 377/2003.

According to Article 158 (1) of the proclamation, before any strike happens, workers must give advance notice to the company explaining in detail the reasons for the strike. As far as I know, no such notices have been submitted to my office. Further, the law says that, for the strikes to be carried out, it should be supported by the majority of workers with a formal meeting. To

my knowledge, no such formal meeting has happened. These two violations by you and your comrades indicate that the current strike is clearly in the contravention of the labor law. If the strikes do not stop before the end of the day, you will hear from our lawyers. We will sue you personally for initiating and implementing the strikes in defiance of the country's law.

Table 3.6 summarises these responses.

Table 3.6: Summary of Managerial Responses to Hypothetical Workplace Conflict

Ranking	Response
Manager 1	**Apologetic, collaborative, empowering** - Apologizes for slow response. - Proposes forming a committee with workers' involvement to address issue. - Pledges resources and seeks cooperation to end strike.
Manager 2	**Acknowledging, proactive, realistic** - Accepts request as fair and promises improvements. - Details immediate and long-term actions needed. - Requests patience and asks for strike to end.
Manager 3	**Appreciative, conciliatory, blame-shifting, behavioral focus** - Thanks representative and recognizes issue. - Blames workers for misusing toilet breaks. - Suggests ending strike and working together to improve behavior before addressing facilities.
Manager 4	**Acknowledging, deflective, postponing, subtle pressure** - Recognizes importance of issue but cites current financial struggles. - Suggests postponing improvements and implies potential career consequences for pushing issue.
Manager 5	**Dismissive, threatening, defensive** - Claims company's facilities are sufficient. - Compares to other companies' and workers' home facilities. - Threatens workers to cease strike or face consequences.
Manager 6	**Legalistic, hostile, threatening, dismissive** - Refuses to discuss issue. - Cites labour law violations and threatens legal action. - Demands strike to end immediately.

Notes This table summarises the managerial responses to the hypothetical workplace conflict, ordered based on the ranking of the senior manager (best to worst). The adjectives describing each of the vignettes were generated by ChatGPT.

3.7.B Experiment

3.7.B.1 Structure of the assessments

To implement the assessments, we first randomly divide our full set of candidates into groups of 15, and our full set of firm managers into groups of 10. We divide these sets of 15 candidates into five *triplets* (labeled 1 through 5) and 10 firms into five *pairs* (labeled A through E). As shown in Table 3.7, each firm pair assesses different triplets of candidates in a rotating sequence across five vignettes:

- In the first vignette, pair A assesses triplet 1, pair B assesses triplet 2, pair C assesses triplet 3, and so on (see the diagonal in Table 3.7).
- Keeping the same triplets and firm pairs, we rotate assignments in subsequent vignettes: for example, in the second vignette, pair B assesses triplet 1, pair C assesses triplet 2, pair D assesses triplet 3, etc.
- This pattern continues until each firm pair has assessed every candidate triplet exactly once.

We implement this procedure for 66 separate groups of 15 candidates, each assessed by a total of 66 groups of eight or ten firms. For the groups of eight firms, for each triplet of candidates one vignette is not assessed.

This rotating design enables us to measure how similarly two HR managers in the same firm pair rank the same candidates within a given vignette (i.e., *within-vignette* agreement) and how these rankings compare across vignettes (i.e., *cross-vignette* agreement). It thus provides a direct way to study consistency in managers' preferences.

Table 3.7: Structure of the Assessments

	Firm Pair				
Candidate Triplet	A	B	C	D	E
1	V1	V2	V3	V4	V5
2	V5	V1	V2	V3	V4
3	V4	V5	V1	V2	V3
4	V3	V4	V5	V1	V2
5	V2	V3	V4	V5	V1

Notes: This table illustrates how one group of 15 candidates is assessed by one group of 10 firms (grouped into five pairs). Each cell shows the vignette (V1–V5) that a particular firm pair assesses for a particular candidate triplet.

3.7.B.2 Matching firms to candidates

We match firms to candidates using a random sequential dictator algorithm. The matching procedure is structured as follows:

1. We divide firms into groups of 10 and, separately, candidates into groups of 15.
2. Each group of candidates is assigned a random preference ordering over all groups of firms.
3. Groups of candidates are matched sequentially based on their preference rankings:
    - Each group of candidates is assigned to their most preferred group of firms that has not yet been matched.
4. We impose an additional constraint: firms cannot assess candidates who have previously interned at their firm, as part of the experiment conducted in (Abebe et al., 2024).

Within each group-level match, triplets of candidates are matched to pairs of firms for the first round of assessments using the same algorithm. After this, triplets of candidates rotate across pairs of firms using the mechanism described in Section 3.7.B.1.

3.7.B.3 A revealed preference exercise with human resource managers

We designed these mechanisms to achieve three objectives: (i) the mechanisms are intuitively simple for respondents to understand, and truthful reporting should be an 'obviously dominant strategy' (within the formal definition of Li (2017)); (ii) as far as possible, the two mechanisms are the same (so that any differences in reporting are attributable to differences in preferences, rather than to the mechanisms themselves), and (iii) we do not mislead the respondents in any way (therefore, any promises that we make in order to incentivise the decision are feasible for us to implement). Our mechanisms are as follows:

1. Suitability as an employed manager: We tell the respondent:

   > *I'm about to ask the computer to randomly choose a number: 1 or 2. This is the number of candidates whose details I will be asked to pass to you. I need you to commit, in advance, to which candidates you would like to see in each circumstance. You have seen three candidates. Suppose that the computer tells me that you may receive the contact details of two. In that situation, of the three candidates, whose details would you then not want to see? Now, suppose that the computer tells me that you may receive the contact details of only one. Of the two candidates remaining, whose details would you then not want to see?"*

   This process continues until just one candidate remains.[27]

2. Suitability as an entrepreneur: We explain to the respondent that we plan to run a series of business plan competitions – in which a set of candidates is judged by experienced business managers, with a winner receiving US$1000 to start (or to support) his or her own business. To elicit a ranking over candidates' entrepreneurial abilities, we nest the previous mechanism in a simple 'random dictator' mechanism. We tell the respondent:

[27] In a follow-up survey with the candidates, twelve report having been contacted by firms, none of these were eventually hired.

> *You are one of two business people who will review this set of three aspiring managers. We will randomly choose just one of the two of you to decide which of the three should be candidates at the business plan competition. Suppose that person is you. The computer will, again, randomly choose a number: 1 or 2. This time, this number will tell us how many candidates will be invited to the business plan competition. You have seen five candidates. Suppose that you are chosen to send candidates to the business plan competition, and that the computer decides that you should send two candidates. In that situation, of the three candidates, who should not be sent to the competition?*

As above, this process continues until just one candidate remains.

These mechanisms are each similar to the 'OSP-RSD' ranking mechanism described in Li (2017); they are simple to understand, and it is an 'obviously dominant' strategy for the respondent to rank truthfully.

3.7.B.4 A stated preference exercise with human resource managers

After each recording, we ask the manager the following questions (each of which uses a five-point Likert scale):

1. Suitability as an employed manager: *Think about the lowest-level managerial position in your firm. Imagine that a firm – either your firm or another firm like yours – is interested in hiring for this position. Based on the recording that you have just heard, how likely are you to recommend hiring this person for a managerial position like that?*

2. Suitability as an entrepreneur: *Suppose that the person you have just seen is acting in the capacity as a founder and managing director of his/her own firm. In your opinion, how likely is it that this person would be a successful entrepreneur?*

These answers record managers' stated preferences about each candidate separately.

3.7.C Details: Encoding of Responses

Table 3.8: Encoding of Responses

Dimension	Enumerator encoding	Encoding for analysis
Action	1 Agrees	1 Agrees
	2 Disagrees	2 Disagrees
Justification	Varies by vignette; see following table	1 Firm's Interest
		2 Other person's interest
		3 Respondent's own interest
		4 Shared Interest
		5 No justification
Authority	1 Formal Authority	1 Formal Authority
	2 Higher Principles	2 Higher Principles
	3 Personal Authority	3 Personal Authority
	4 Personal Relationship*	4 No Source
	5 No Source of Authority	
Tone		1 Aggressive
		2 Assertive
		3 Calm/Assured
		4 Timid

* This is almost never used.

Table 3.9: Encoding Justification for Various Scenarios

Encoding Enumerator	Encoding for analysis
Line Management of an Employee	
The action is to protect/help the respondent's company	1. Firm's Interest
The action is in the interests of the other employees.]	
The action is to protect/help the employee	2. Other party's interest
The action is in the respondent's own interest	3. Respondent's own interest
The respondent did not justify his or her action	5. No justification
Negotiating with a Supplier	
The action is to protect/help the respondent's company	1. Firm's Interest
The action is to protect/help the supplier	2. Other party's interest
The action is in the respondent's own interest	3. Respondent's own interest
Reference to the relationship between the respondent's firm and the supplier	4. Shared interest
The respondent did not justify his or her action	5. No justification
Other*	
Negotiating a Pay Rise	
The action is to protect/help the respondent's company	1. Firm's Interest
The action is to protect/help the employee	2. Other party's interest
The action is in the respondent's own interest	3. Respondent's own interest
Reference to the relation with the employee and his/her family	4. Shared interest
Reference to the employee's good performance	
Reference to the effect it might have on other employees	1. Firm's Interest
The respondent did not justify his or her action	5. No justification
Other*	5. No justification
Negotiate an Adjustment with the Banks	
The action is to protect/help the respondent's company	1. Firm's Interest
The action is to protect/help the bank	2. Other party's interest
The action is in the respondent's own interest	3. Respondent's own interest
Reference to the relationship between the firm and the bank	4. Shared interest
Reference to the inability of the company to repay the loan	1. Firm's Interest
The respondent did not justify his or her action	5. No Justification
Other*	5. No justification
Negotiate with a Client	
The action is to protect/help the respondent's company	1. Firm's Interest
The action is to protect/help the client	2. Other party's interest
The action is in the respondent's own interest	3. Respondent's own interest
Reference to the relationship between the respondent's firm and the client	4. Shared interest
The respondent did not justify his or her action	5. No justification
Other*	5. No justification

* We manually determined "Other" implies no justification based on the enumerators' open-ended responses.

3.7.D Further detail on intentions

Table 3.10: Reported intentions of types by vignette

Vignette	Type	Expresses distrust counterparty	Maintain relationship	Follow procedure	Set an example
Line Management	Authoritative	.58	0	.51	.38
	Affiliative	.41	0	.39	.32
	Coercive	.49	0	.36	.28
	Timid	.42	0	.44	.28
Supplier	Authoritative	.03	0	.30	.53
	Affiliative	.03	.02	.21	.44
	Coercive	.03	0	.21	.46
	Timid	.04	0	.16	.46
Pay rise	Authoritative	0	.12	.61	.34
	Affiliative	0	.21	.28	.24
	Coercive	0	.08	.50	.18
	Timid	0	.06	.47	.20
Bank	Authoritative	0	.71	0	0
	Affiliative	0	.75	0	0
	Coercive	0	.65	0	0
	Timid	0	.59	0	0
Client	Authoritative	.61	.20	.45	.07
	Affiliative	.23	.50	.18	.02
	Coercive	.60	.21	.35	.02
	Timid	.55	.19	.29	.02

Notes: This table describes the intentions of respondents' for each vignette, which are enumerated before the respondent actually responds to the vignette. The four characteristics are (i) whether the respondent expresses they do not trust the other part, (ii) whether the respondent wishes to maintain a good relationship, (iii) whether the respondent mentions following procedure, and (iv) whether the respondent wishes to set an example for future interactions. These numbers are calculated by assigning each individual to the pure type for which they have the highest estimated $\hat{\boldsymbol{\theta}}_i$. Then, a conditional average is taken for each pure type.

3.7.E Bayesian modelling in Stan

3.7.E.1 Dirichlet model

We observe individuals $i \in \{1, \ldots, N\}$ performing on vignettes $v \in \{1, \ldots, 5\}$. For each individual assessment of a vignette, we have a set of two enumerators, $e \in \{1, 2\}$. Each enumerator records a set of 'attributes' of the response (action, authority, justification and tone): $a \in \{1, \ldots, 4\}$. Each attribute a has $J(a)$ possible categorical responses.[28] $y^a_{ive} \in \{1, \ldots, J(a)\}$ is the response recorded by enumerator e for vignette v for individual i for attribute a.

For attribute a, consider a 'pure type' $k \in \{1, \ldots, K\}$, having a vector $\boldsymbol{\phi}_{ka}$ of dimension $J(a)$, such that $\boldsymbol{\phi}_{ka} \in (0, 1)^{J(a)}$. Denote by $\tilde{\boldsymbol{\phi}}_{ka}$ the inverse Multinomial Logit transformation of $\boldsymbol{\phi}_{ka}$.[29] For simplicity and tractability, we assume that, conditional on $\boldsymbol{\phi}_{ka}$, the attributes are realised independently as:

$$y^a_{ive} \,|\, \boldsymbol{\phi}_{ka}, \boldsymbol{\psi}_{av}, \boldsymbol{\chi}_{ev} \sim_{iid} \texttt{Multinomial Logit}\left(\tilde{\boldsymbol{\phi}}_{ka} + \boldsymbol{\psi}_{av} + \boldsymbol{\chi}_{ae}\right), \tag{3.12}$$

where $\boldsymbol{\psi}_{av}$ and $\boldsymbol{\chi}_{ae}$ here represent $J(a)$-element vectors.[30][31]

This is essentially a Latent Dirichlet Allocation model, with a few adjustments for our setting:

[28] Specifically, $J(1) = 2$ for the action; $J(2) = 4$ for source of authority, $J(3) = 5$ for justification and $J(4) = 3$ for tone.

[29] That is, for the sth element of $\boldsymbol{\phi}_{ka}$ – which we denote $\boldsymbol{\phi}^{(s)}_{ka}$ – we have:

$$\boldsymbol{\phi}^{(s)}_{ka} = \frac{\exp\left(\tilde{\boldsymbol{\phi}}^{(s)}_{ka}\right)}{1 + \sum_{m=1}^{J(a)} \exp\left(\tilde{\boldsymbol{\phi}}^{(t)}_{ka}\right)} \iff \tilde{\boldsymbol{\phi}}^{(s)}_{ka} = \ln\left(\frac{\boldsymbol{\phi}^{(s)}_{ka}}{1 - \sum_{m=1}^{J(a)} \boldsymbol{\phi}^{(t)}_{ka}}\right).$$

[30] Where the first element of these vectors are normalised to zero, and the whole vector is normalised to zero for the first respectively vignette and enumerator.

[31] $\boldsymbol{\psi}_{av}$ and $\boldsymbol{\chi}_{ae}$ provide a measurement scheme; this is similar in spirit to Item Response Theory. That is, this captures the notion that different enumerators will have different underlying tendencies to assess different attributes, and similarly that for different vignettes respondents have different underlying tendencies to display traits. Note that, in the special case that $\boldsymbol{\psi}_{av} = \mathbf{0}$ and $\boldsymbol{\chi}_{ae} = \mathbf{0}$, this collapses to a more standard Latent Dirichlet Allocation.

1. $\boldsymbol{\psi}_{av}$ and $\boldsymbol{\chi}_{ae}$ provide a measurement scheme; this is similar in spirit to Item Response Theory. (That is, this captures the notion that different enumerators will have different underlying tendencies to assess different attributes on different vignettes.) Note that, in the special case that $\boldsymbol{\psi}_{av} = \mathbf{0}$ and $\boldsymbol{\chi}_{ae} = \mathbf{0}$, this collapses to a more standard Latent Dirichlet Allocation.

2. By having independent 'attributes', we reduce the state space.

3. Each attribute has a different "dictionary".

4. For each vector $\boldsymbol{\psi}_{ka}$ and $\boldsymbol{\chi}_{ea}$ the first element is normalised to zero.

We can then write the following generative model:

1. Draw $\boldsymbol{\phi}_{ka}$ independently for $k \in \{1, \ldots, K\}$ and for $a \in \{1, \ldots, A\}$ from $\texttt{Dirichlet}(\zeta)$;

2. Draw $\boldsymbol{\theta}_i$ independently for $i \in \{1, \ldots, N\}$ from $\texttt{Dirichlet}(\eta)$, where $\boldsymbol{\theta}_i$ is a K-dimensional vector;

3. Draw $\boldsymbol{\psi}_{av}$ and $\boldsymbol{\chi}_{ev}$ from some suitable prior (respectively $\mathcal{N}(0, \sigma_\psi^2)$ and $\mathcal{N}(0, \sigma_\chi^2)$);

4. For each vignette, draw a type z_{iv} from $\boldsymbol{\theta}_i$ and draw an attribute from $y_{aive} \mid \boldsymbol{\phi}_{z_{iv},a}, \boldsymbol{\psi}_{av}, \boldsymbol{\chi}_{ae}$ (using the Multinomial Logit formula above).

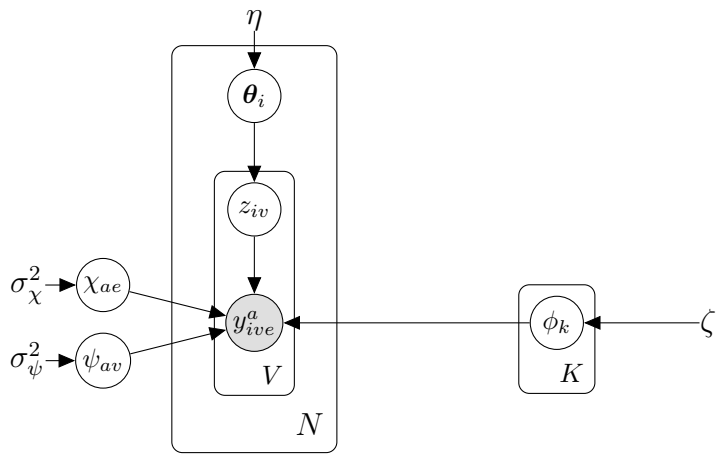
This model can described graphically as follows:

The total probability of the model is as follows, denoting by W^q the qth attribute within a vignette and by β^q the set of parameters for that attribute:[32]

$$P(W, Z, \boldsymbol{\theta}, \boldsymbol{\phi}, \eta, \zeta) = \prod_{j=1}^{K} P(\boldsymbol{\phi_j}; \zeta) \prod_{i=1}^{N} P(\boldsymbol{\theta_i}; \eta);$$
$$\prod_{v=1}^{V} \sum_{k=1}^{K} \left(P(z_{iv} = k | \boldsymbol{\theta_i}) \prod_{e=1}^{2} \prod_{a=1}^{A} P(y_{ive}^a | \boldsymbol{\phi}_{ka}, \boldsymbol{\psi}_{av}, \boldsymbol{\chi}_{ae}) \right),$$

[32] For clarify of exposition we omit the normal prior for ψ and χ from this equation. These enter the log-likelihood linearly.

Figure 3.6: Augmented Latent Dirichlet Allocation: Plate diagram



where $P(\boldsymbol{\phi_j}; \zeta)$ and $P(\boldsymbol{\theta_i}; \eta)$ follow a latent Dirichlet distribution, $P(z_{iv}|\theta_i)$ follows a categorical distribution and $P(y^q_{ive}|\boldsymbol{\phi}_{z_{iv}}, \boldsymbol{\psi}_{av}, \boldsymbol{\chi}_{ae})$ follow multinomial logit distributions.

To fully specify the model, we need to impose an additional set of priors; we choose:

$$\alpha = 0.3;$$

$$\beta = 1;$$

$$\boldsymbol{\chi}_{ae} \sim \mathcal{N}(0, 3);$$

$$\boldsymbol{\psi}_{av} \sim \mathcal{N}(0, 3).$$

In practice, the model is highly insensitive to the choice of the prior β, as the data is highly informative for the parameter $\boldsymbol{\phi}_k$. However, the model is more sensitive to the choice of the parameter α since the parameter θ_i is estimated based only on data from a single candidate. The primary effect of the choice of α is the distance to the edge of the simplex of resulting parameter estimates.

3.7.E.2 Plackett-Luce model

To analyse the ranking data from the HR managers we develop the following Plackett-Luce model. The model employs a Bayesian hierarchical structure where preferences for types and attributes can vary across firms. We include an individual fixed effect for each individual to improve the precision of the resulting estimates:

$$U_{fiv} = \boldsymbol{\beta}_f \boldsymbol{\theta}_i + \boldsymbol{\alpha}_f x_i + \gamma_i + \varepsilon_{fiv}, \tag{3.13}$$

where we have the following explanatory variables:

$$\begin{aligned} \boldsymbol{\theta}_i \implies & \quad \text{an estimated parameter from the first-stage model; } \tilde{\boldsymbol{\theta}}_i \\ x_i \implies & \quad \text{“Known”.} \end{aligned}$$

We specify the error term as follows:

$$\varepsilon_{fiv} \sim EV1$$

Then we employ a Bayesian hierarchical structure for each set of parameters. For the individual mean γ_i, where we specify the error structure for the variance on a bounded uniform distribution; we specify:[33]

$$\begin{aligned} \gamma_i &\sim \mathcal{N}(0, \sigma_\gamma); \\ \sigma &\sim \mathrm{U}(0, 10). \end{aligned}$$

For identification, we set $\gamma_i = 0$ for the first member of every triplet. Then for the $\boldsymbol{\alpha}$

[33] In setting the prior for the variance as a uniform distribution we follow Gelman (2006).

and $\boldsymbol{\beta}$ parameters we use the following structure:

$$\boldsymbol{\alpha}_f \sim \mathcal{N}(\boldsymbol{\alpha}, \sigma_\alpha^2);$$
$$\boldsymbol{\alpha} \sim \mathcal{N}(0, 3);$$
$$\sigma_\alpha^2 \sim \mathrm{U}(0, 10).$$

and

$$\boldsymbol{\beta}_f \sim \mathcal{N}(\boldsymbol{\beta}, \sigma_\beta^2);$$
$$\boldsymbol{\beta} \sim \mathcal{N}(0, 3);$$
$$\sigma_\beta^2 \sim \mathrm{U}(0, 10).$$

This implies the following total probability (denoting by θ the parameters, by y the data, and by λ the hyperparameters, and denoting $U(\boldsymbol{\beta}_f, \boldsymbol{\alpha}_f, \gamma_i, \boldsymbol{\theta_i}, \boldsymbol{x}_i) \equiv U_{fiv}$):

$$P(y|\boldsymbol{\beta}_f, \boldsymbol{\alpha}_f, \gamma_i, \boldsymbol{\theta_i}, \boldsymbol{x}_i) = \prod_{f=1}^{F} \prod_{v=1}^{5} \frac{\exp\left(U_{fiv}^1\right)}{\sum_{r=1}^{3} \exp\left(U_{fiv}^r\right)} \cdot \frac{\exp\left(U_{fiv}^2\right)}{\sum_{r=2}^{3} \exp\left(U_{fiv}^r\right)}.$$

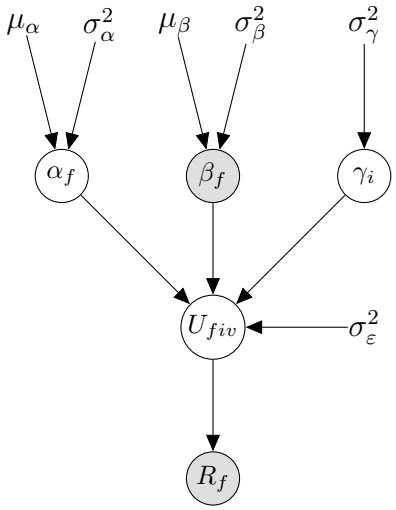
We then specify the distribution of α_f, γ_i and β_f as:

$$P(\boldsymbol{\beta}_f, \boldsymbol{\alpha}_f, \gamma_i|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\gamma, \sigma_\alpha, \sigma_\beta) = \prod_{i=1}^{N} (\phi(\boldsymbol{\gamma}_i|0, \Sigma_\gamma) \prod_{f=1}^{F} \phi(\boldsymbol{\alpha}_f|\boldsymbol{\alpha}, \sigma_\alpha) \cdot \phi(\boldsymbol{\beta}_f|\boldsymbol{\beta}, \sigma_\beta).$$

Finally, we define the hyperprior as:

$$P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\gamma, \sigma_\alpha, \sigma_\beta) = \phi(\boldsymbol{\alpha}|0, 10) \cdot \phi(\boldsymbol{\beta}|0, 10) \cdot U(\sigma_\alpha, 0, 10) \cdot U(\sigma_\beta, 0, 10) \cdot U(\sigma_\gamma, 0, 10).$$

The full probability of drawing a set of parameters is the product of these three elements (following Bayes' rule):



3.7.E.3 Estimation

We estimating using Hamiltonian Monte Carlo (HMC), implemented in Stan (Carpenter, Gelman, Hoffman, Lee, Goodrich, Betancourt, Brubaker, Guo, Li, and Riddell, 2017; Gelman, Lee, and Guo, 2015; Neal and Neal, 1996). HMC is an advanced algorithm for Markov Chain Monte Carlo (MCMC) simulations. Its key feature is the use of the gradient of the posterior distribution in determining step sizes. This allows HMC to consider how the probability of the parameters changes, taking more informed steps during the simulation. This method is particularly useful because it often converges to high-probability regions faster and more efficiently than traditional MCMC methods. This is especially true in complex models with many parameters, like the one we implement. This efficiency is due to its ability to avoid getting stuck in less probable regions, a common issue with simpler MCMC methods.

3.7.E.4 The Dirichlet model

We first estimate the Dirichlet model, and use these results for Section 3.3. A few things are worth noting: Firstly, the commonly used Gibbs Sampler (Steyvers, Smyth, Rosen-Zvi, and Griffiths, 2004) is not applicable due to the additional parameters in our model. However, the small number of types and attributes makes a direct estimation procedure feasible. Secondly, a K-component mixture distribution is invariant to permutations of

Table 3.11: Convergence statistics for the Dirichlet model with four types

Parameter	Average PSRF	Highest PSRF	Average effective sample size	Lowest effective sample size
Theta	0.999	1.009	2000	247
Phi Action	1.010	1.011	738	661
Phi Tone	0.999	1.001	1980	1359
Phi Justification	1.001	1.004	1462	965
Phi Authority	1.002	1.006	1116	791
Chi Action	0.999	0.999	2000	2000
Chi Tone	0.998	0.998	2000	2000
Chi Justification	0.999	1.000	2000	1748
Chi Authority	1.001	1.001	2000	1863
Psi Action	1.008	1.013	831	676
Psi Tone	0.999	1.002	2000	2000
Psi Justification	0.999	1.003	1777	1043
Psi Authority	1.001	1.006	1311	1228

Notes: Table 3.11 shows the average and highest PSRF, and the average and lowest effective sample size for each subset of the parameters in the Dirichlet model. Crucially, the highest PSRF is low, showing good convergence of the model.

component labels, resulting in $K!$ modes in the posterior distribution of the mixture parameters. This can be problematic when using HMC, as the leapfrog estimator may attempt to jump between different configurations of the component labels. In practice, the distinct modes of our posterior distribution prevent label switching in our draws. Consequently, we can re-label the types for consistent labeling across chains before performing standard convergence tests for MCMC simulations (Gelman and Rubin, 1992). To make inference easier, in particular for the joint model, we first run four shorter chains which we then relabel to each have the same type labels. We then use the estimates from these four shorter chains to start four new chains, which we use – following burn-in – in our analysis.

3.7.E.5 The Placket-Luce model and joint-estimation

We estimate the Plackett-Luce model using Hamiltonian Monte Carlo (HMC) in STAN, implementing a non-centered reparameterization of the hierarchical parameters. This technique results in faster convergence and increases the precision of the model estimates. In such a nonlinear model, incorporating uncertainty in the first-stage parameters—the

parameters of the Latent Dirichlet Allocation (LDA) model—presents a challenge. Properly propagating this uncertainty to the second-stage model provides a more realistic representation of the model's confidence in its predictions and estimates. To address this, we follow Battaglia et al. (2024), jointly estimating the LDA and Plackett-Luce models. This joint estimation approach leverages the curvature of the log-likelihood to guide the level of uncertainty about the type parameters, implicitly weighting the observations in the second-stage model by the uncertainty about θ_i.

Tables 3.12 and 3.13 show the potential scale reduction factor and average effective sample size of the chains. This is based on four chains with 1000 draws.

Table 3.12: Convergence statistics for the joint model with four types for ranking as entry-level manager

Parameter	Average PSRF	Highest PSRF	Average effective sample size	Lowest effective sample size
Theta	1.000	1.007	4000	409
Phi Action	1.006	1.007	845	777
Phi Tone	1.000	1.002	3223	1891
Phi Justification	1.007	1.012	583	358
Phi Authority	1.001	1.004	1937	1298
Chi Action	0.999	0.999	4000	4000
Chi Tone	0.999	1	3496	3431
Chi Justification	0.999	1	4000	4000
Chi Authority	0.999	0.999	3268	3216
Psi Action	1.005	1.007	1013	870
Psi Tone	1.001	1.001	3220	2595
Psi Justification	1.000	1.011	2626	379
Psi Authority	1.000	1.003	1987	1776
Beta	1.004	1.004	1115	1066
Beta FS	0.999	1.006	4000	771
Beta Std	1.028	1.054	352	180
Gamma	0.999	1.001	4000	1349
Gamma Std	1.001	1.001	1232	1232

Notes: This table reports the average and highest PSRF, and the average and lowest effective sample size for each subset of the parameters in the jointly estimated model for the rankings as potential entry-level managers. Crucially, the highest PSRF is low, showing good convergence of the model.

Table 3.13: Convergence statistics for the joint model with four types for ranking as entrepreneur

Parameter	Average PSRF	Highest PSRF	Average effective sample size	Lowest effective sample size
Theta	1.000	1.008	4000	387
Phi Action	1.006	1.008	1022	943
Phi Tone	1.001	1.004	2449	1475
Phi Justification	1.018	1.027	419	241
Phi Authority	1.002	1.007	1746	993
Chi Action	0.999	0.999	4000	4000
Chi Tone	1.002	1.003	2709	2662
Chi Justification	1.000	1.001	4000	4000
Chi Authority	1.000	1.000	2821	2814
Psi Action	1.005	1.009	1080	947
Psi Tone	1.002	1.004	2349	1960
Psi Justification	1.001	1.027	2619	243
Psi Authority	1.003	1.006	1528	1262
Beta	1.006	1.006	1124	1075
Beta FS	0.999	1.003	4000	672
Beta Std	1.023	1.027	252	182
Gamma	0.999	1.002	4000	846
Gamma Std	1.002	1.002	1285	1285

Notes: This table reports the average and highest PSRF, and the average and lowest effective sample size for each subset of the parameters in the jointly estimated model for the rankings as potential entrepreneurs. Crucially, the highest PSRF is low, showing good convergence of the model.

3.7.E.6 Stability of type parameters with joint estimation

Figures 3.7 and 3.8 show that the estimated type probability and distribution over type is consistent in the individually estimated Dirichlet model and the jointly estimated model.

Figure 3.7: The type parameters and distribution over types in the Dirichlet model

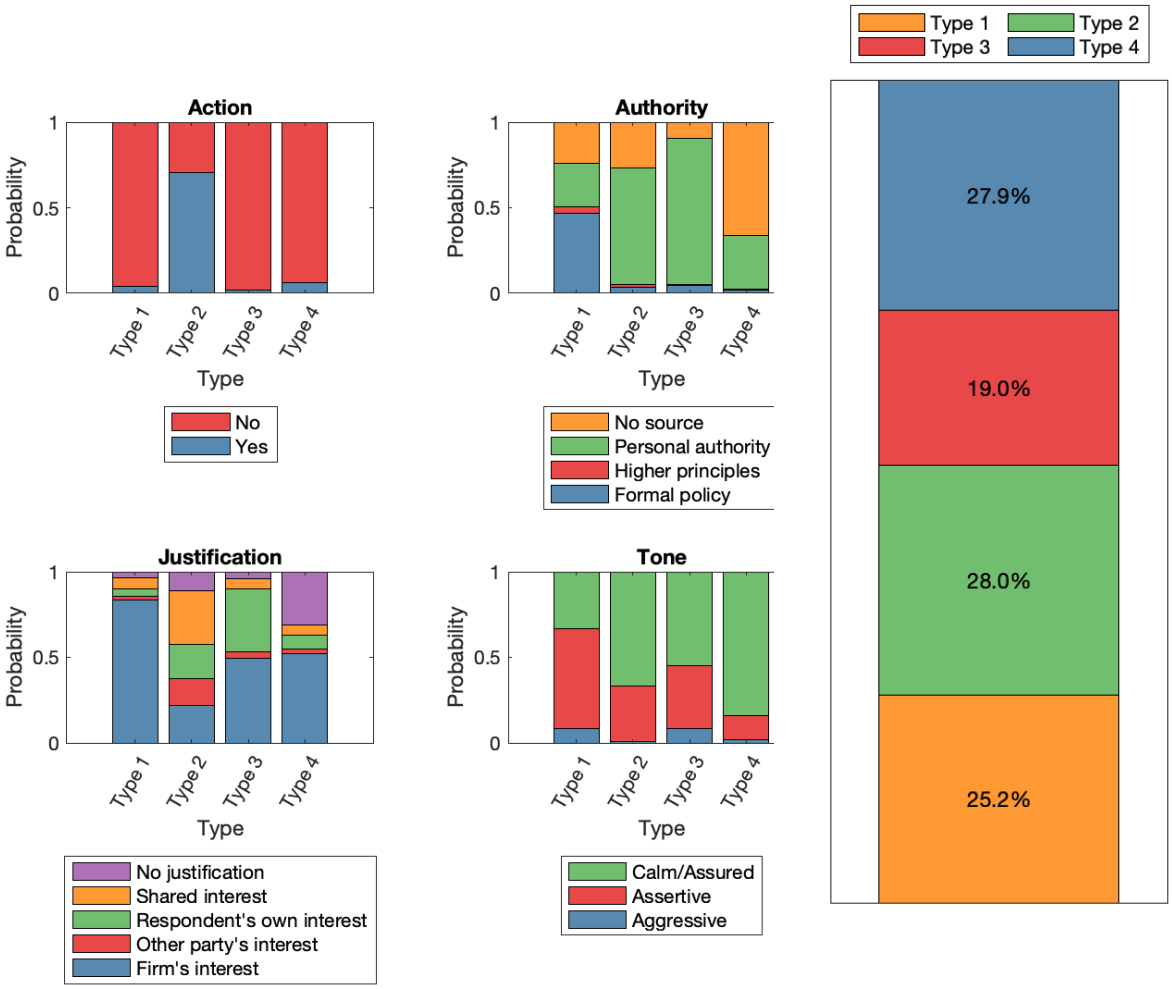
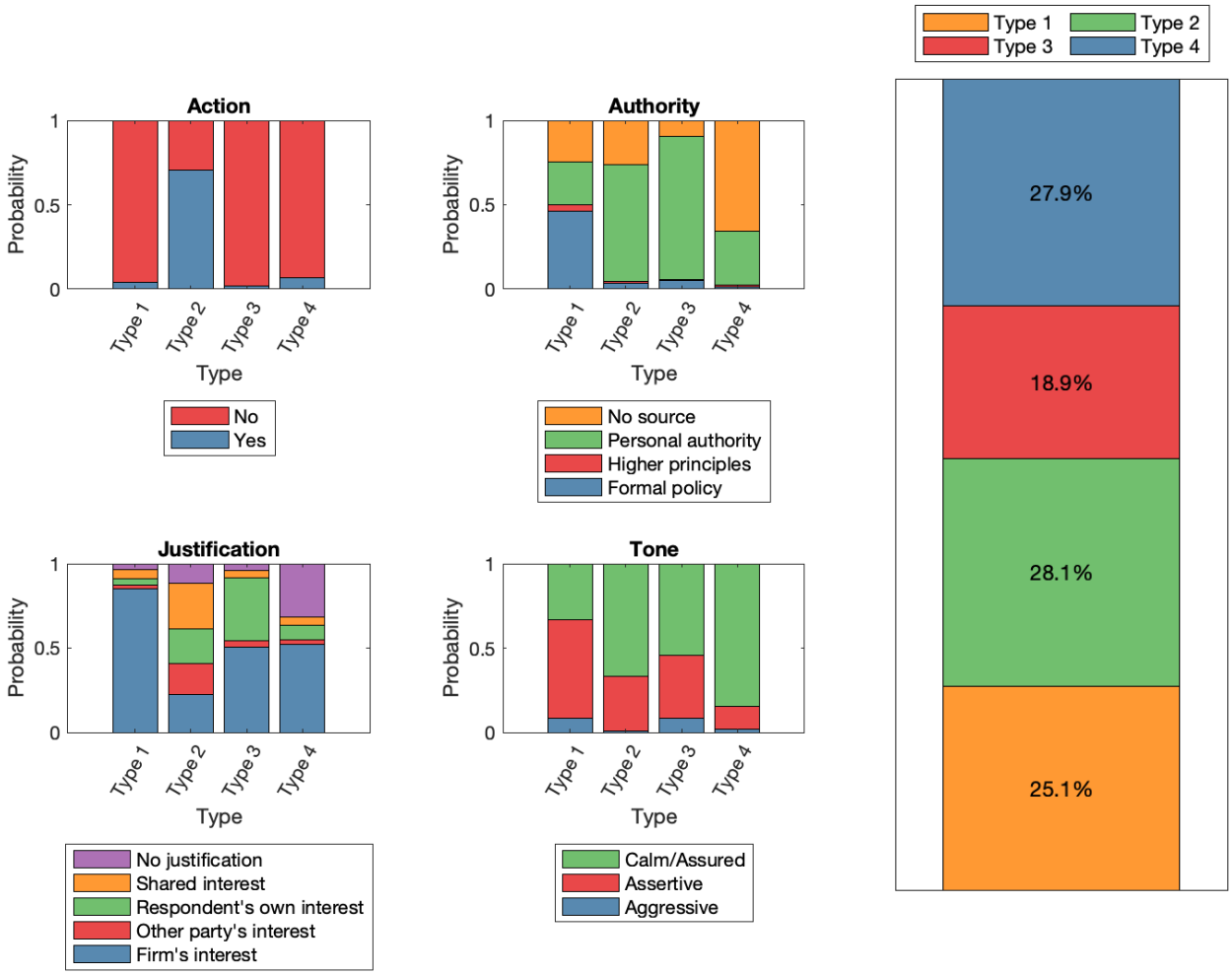


Figure 3.8: The type parameters and distribution over types in the joint model



3.7.F Sensitivity to the choice of number of types

Our preferred estimation uses $K = 4$ types. In this appendix section, we consider alternative specifications with two, three and five types. In short, we argue that (i) our results are robust to using $K = 2$ or $K = 3$, and (ii) the model fit becomes unstable for $K = 5$ (in a manner that we explain shortly), and the model with $K = 5$ does not meaningfully improve on the fit compared to $K = 4$.

3.7.F.1 Considering two types and three types

On the following pages, we repeat each of the key estimations from our preferred model, but using $K = 2$ and then $K = 3$. By way of summary, the following key results of the paper are replicated across these two alternative specification:

1. In terms of attributes of the responses, we find an 'affiliative'-like and 'authoritative'-like type in all specifications. For $K = 2$, these are essentially the two types that we find. For $K = 3$, we have an affiliative type, a authoritative type and a coercive/timid type. When $K = 4$, the latter two types are split into an authoritative, covercive and timid type. We illustrate this in Figure 3.9. This figure shows the key intuition for why our results are stable across the choice of types: each specification preserves the essential distinction between the authoritative type and the other type(s).

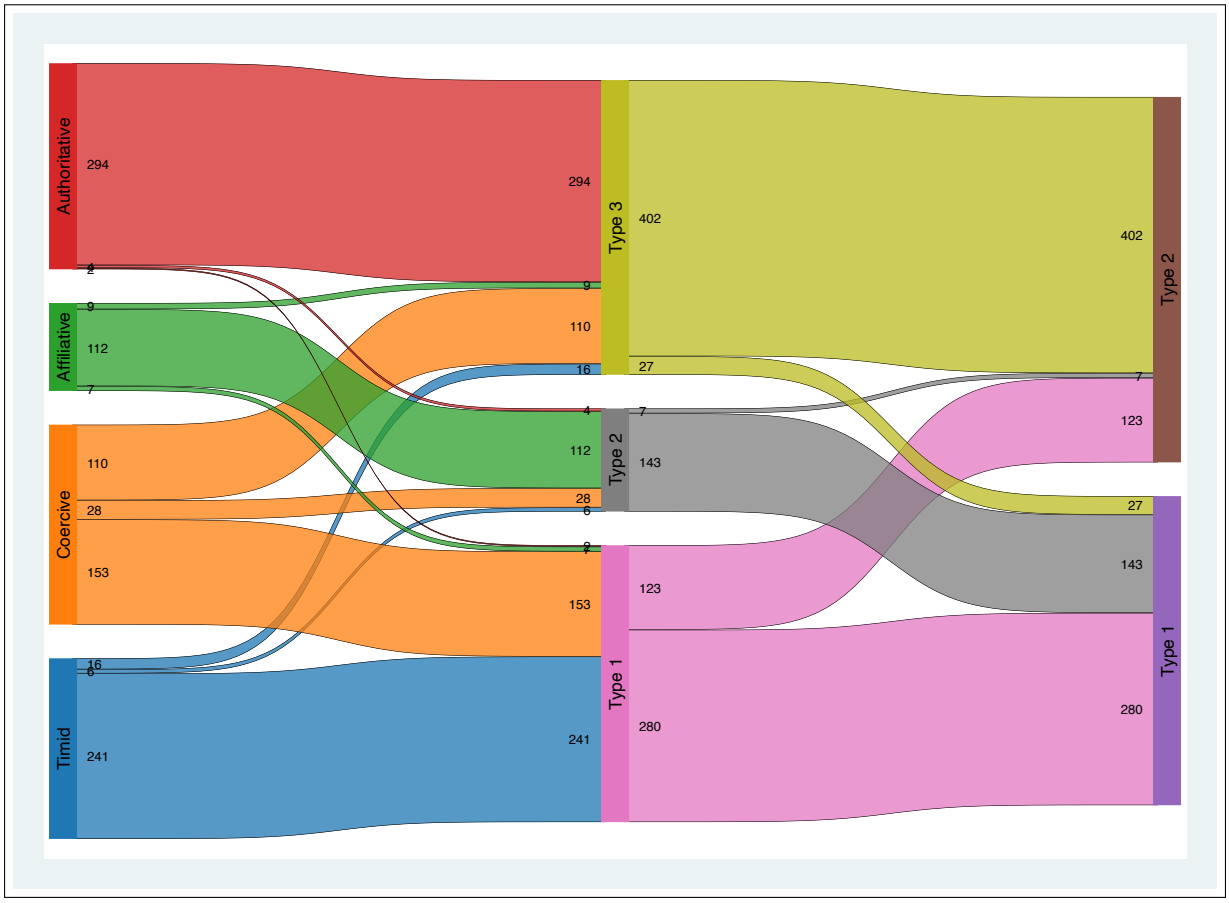
2. For $K = 2$, $K = 3$ and $K = 4$, the authoritative type is in each case more likely to be male, has better labour market outcomes, and is better able to predict the decision of the senior manager.

3. For $K = 2$, $K = 3$ and $K = 4$, the management experience treatment causes candidates with low parental education to be more likely to be authoritative. For $K = 3$ and $K = 4$, we learn that this is driven by a shift from a coercive to an authoritative management style (an insight that is not possible for the $K = 2$ model).

4. For $K = 2$, $K = 3$ and $K = 4$, the authoritative type is strongly preferred by firm managers.

5. The effect of the first actors' gender on the realised type is comparable across specifications, but most pronounced when $K = 4$.

Figure 3.9: Sankey flow diagram: Assignment of types under $K = 2$, $K = 3$ and $K = 4$



3.7.F.2 Results: Two-type model

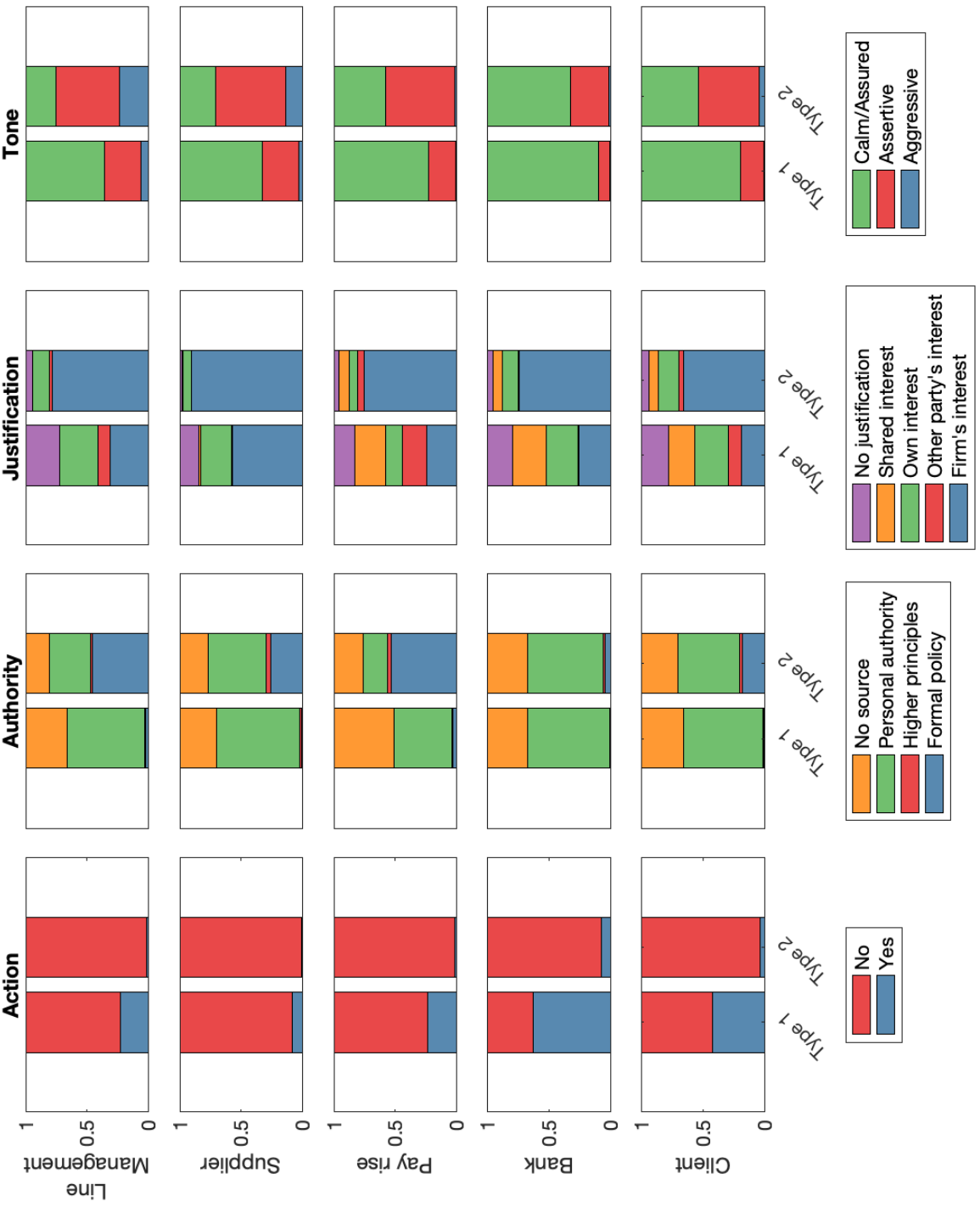


Figure 3.10: 'pure type' management styles amongst Ethiopian young professionals (2-type model)

Figure 3.11: Distribution of types across individuals (2-type model)

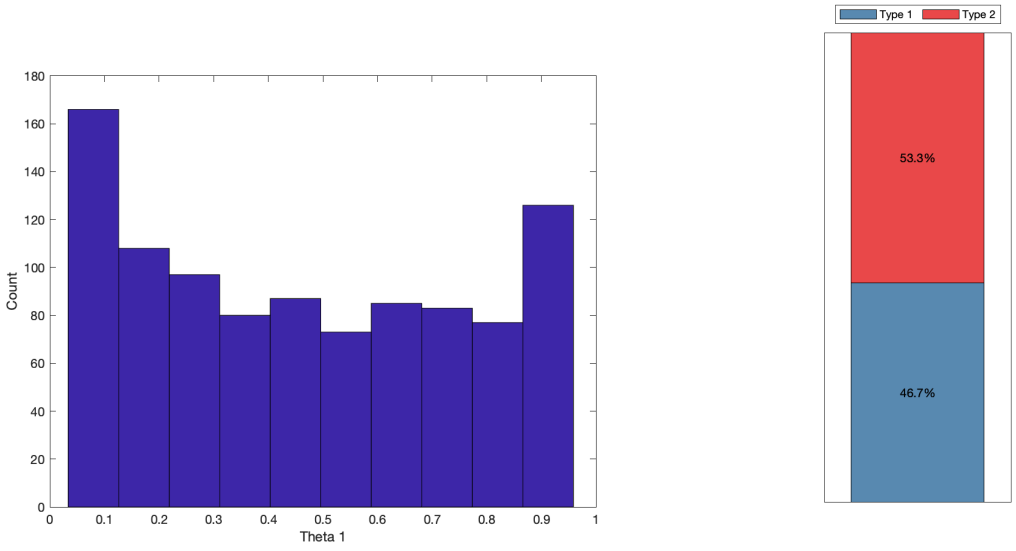


Table 3.14: Agreement in rankings between the respondents' and the HR consultant's rankings. (2-type model)

	Prediction Senior manager and senior manager	Own action and Senior Manager	Internal Agreement
Overall	.511	.415	.793
Type 1	.464	.382	.814
Type 2	.551	.443	.775

Notes This table displays Kendall's Tau values comparing the respondents' rankings and the expert's rankings. The column "Prediction of Senior Manager versus Actual Senior Manager" compares the respondents' predictions of the senior manager's actions with the actual senior manager's actions. The column "Own Action versus the Senior Manager" compares the senior manager's actions with what the respondent would do themselves. The column "Internal Agreement" displays Kendall's Tau for the respondents' perception of their own actions versus what they expect a senior manager would do. Kendall's Tau is calculated as $\tau = \frac{\sum \text{agreements} - \sum \text{disagreements}}{\sum \text{pairwise comparisons}}$.

Table 3.15: Characteristics and Types: Summary Statistics (2-type model)

Type	Gender [1=male]	Wage employment indicator	Self-Employment indicator	Above Median Reservation Wage	Above Median Reservation Profit	Average Duration Response
Type 1	.781	.668	.149	.287	.300	36.0
Type 2	.818	.699	.174	.404	.408	50.5
p-value	.038	.086	.29	<0.001	<0.001	<0.001

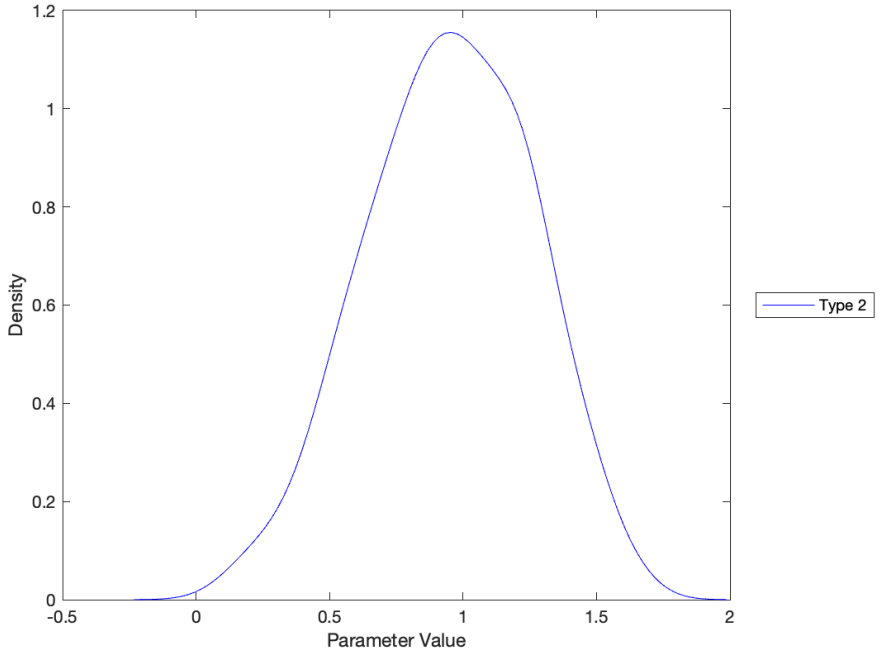
Notes This table describes the average characteristics of individuals of each type. Specific, this includes their gender (1 indicating male, 0 female), a dummy for their wage- and self-employment status, and the probability they have an above-median reservation wage and profit based on data collected before respondents' attended the studio. We also include a dummy for whether or not the individual was treated in the management placement experiment, and the average duration of the responses of the candidate across the vignettes. Note that the median splits on reservation wage and profit do not yield a 50/50 split due to bunching in the underlying data at ETB10.000. These numbers are calculated by assigning each individual to the pure type for which they have the highest estimated $\hat{\boldsymbol{\theta}}_i$. Then, a conditional average is taken for each pure type. To test whether intentions differ across types, we compute the Mahalanobis (Wald) statistic and compare it to a χ^2_3 distribution. This relies on the multivariate normal approximation to the posterior described in Section 4.1 of Gelman et al. (2013). The resulting p-value appears in the final row. For non-binary outcomes, we first create a binary split at the median to split the sample.

Table 3.16: Labour market experience and Types: Summary Statistics (2-type model)

Type	Employment	Permanent employment	Unemployed	Management Position	Number of transitions
	Years	*Years*	*Years*	*Share*	*Count*
Type 1	5.217	3.564	.525	.117	.942
Type 2	5.417	3.801	.417	.168	.735
p-value	.058	.009	.022	.002	.006

Notes This table describes the labour market experience over the past 6 years of individuals of each type. Specific, this includes the number of years they have been in employment including both self- and wage-employment, the number of years they have been in permanent employment, the number of years they have been unemployed, whether they were in a management position before participating in the studio experiment, and finally the number of labour market transitions they have gone through. These numbers are calculated by assigning each individual to the pure type for which they have the highest estimated $\hat{\boldsymbol{\theta}}_i$. Then, a conditional average is taken for each pure type. To test whether intentions differ across types, we compute the Mahalanobis (Wald) statistic and compare it to a χ^2_3 distribution. This relies on the multivariate normal approximation to the posterior described in Section 4.1 of Gelman et al. (2013). The resulting p-value appears in the final row. For non-binary outcomes, we first create a binary split at the median to split the sample.

Figure 3.12: The distribution of the preferences for entry-level managers (2-type model)



Notes This figure shows the distribution of the estimates for $\hat{\boldsymbol{\beta}}_{fs}$ across all firms in terms of demand for entry-level managers. In this figure, the Type 1 is omitted and the estimates for $\hat{\beta_{\text{Type 2}}}$, $\hat{\beta_{\text{Type 3}}}$ are plotted

Table 3.17: The causal effect of managerial experience on management style by parents education (2-type model)

	Full sample		Low parental education		High parental education	
	(1)	(2)	(3)	(4)	(5)	(6)
Type 1 (%)	47.9	-2.2	52.3	-4.5	43.9	0.0
		[-4.7, 1.0]		[-8.3, -0.7]		[-3.5, 3.4]
Type 2 (%)	52.1	2.2	47.7	4.5	56.1	-0.0
		[-1.0, 4.7]		[0.7, 8.3]		[-3.4, 3.5]
N	479	500	229	239	250	261

Notes This table reports the treatment effect of the management experience experiment on the managerial traits of individuals. The treatment effect is calculated based on the distribution of the difference in the average value of θ for treated and untreated individuals. Columns (1), (3) and (5) report the average estimated value of $\boldsymbol{\theta}_i$ for individuals that were not treated in the management experience experiment for respectively all individuals, individuals whose parents did not finish primary school and for individuals for whom at least one parent did. Columns (2), (4) and (6) report the treatment effect of the management experience experiment on their managerial traits for these three groups respectively. In columns (2), (4) and (6) both the average treatment effect and the 95% credible interval, in square brackets, are reported.

Table 3.18: Actors' gender and management styles (2-type model)

	Panel A: Effect on estimated types	
	Theta 1	Theta 2
First actor female	-0.022	0.022
Constant	0.478***	0.522***
Bayesian Credible Interval	[-.046 .004]	[-.004 .046]
N	982	982

	Panel B: Effect on attributes			
	Justification		Authority	
	Rely on firm's interest	Rely on own interest	Rely on formal authority	Rely on seniority
First actor female	0.032*	-0.028**	0.021	-0.026
	(0.019)	(0.014)	(0.015)	(0.020)
Constant	0.554***	0.178***	0.154***	0.537***
	(0.013)	(0.010)	(0.010)	(0.014)
Enumerator FE	Yes	Yes	Yes	Yes
Vignette FE	Yes	Yes	Yes	Yes
Mean dep. var	0.552	0.168	0.163	0.511
N	6887	6887	6887	6887

	Panel C: Effect on managers' assessments			
	Ranking Data		Normalised likert score	
	Manager	Entrepreneur	Manager	Entrepreneur
First actor female	0.052**	0.060***	0.083**	0.093***
	[0.010 0.093]	[0.017 0.102]	(0.033)	(0.033)
Constant	0.474***	0.470***	-0.045*	-0.050**
	[0.453 0.493]	[0.449 0.492]	(0.025)	(0.025)
Vignette FE			Yes	Yes
N			6874	6869

Notes: This figures displays the causal link between the first actor a respondent sees and their subsequent responses (for the second to fifth vignette). Panel A shows that this indeed manifests as a reliance on a more authoritative management style (captured approximately by type 2), reporting 95% Bayesian Credible Intervals in square brackets. Panel B shows that respondents act more like an authoritative type - they agree less, rely on formal policy and authority - after starting with a female actor. These attributes are selected to illustrate the shift from a coercive to an authoritative management style; Appendix 3.7.L reports a sequence of multinomial logit regressions on the full set of attributes from which these were selected. The third panel shows the effect of (a) the first actor being female and (2-3) the results from an Acharya-style mediation analysis for using first the estimates for the type parameter from Panel B) and then the attributes of the responses from Panel A. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

3.7.F.3 Results: Three-type model

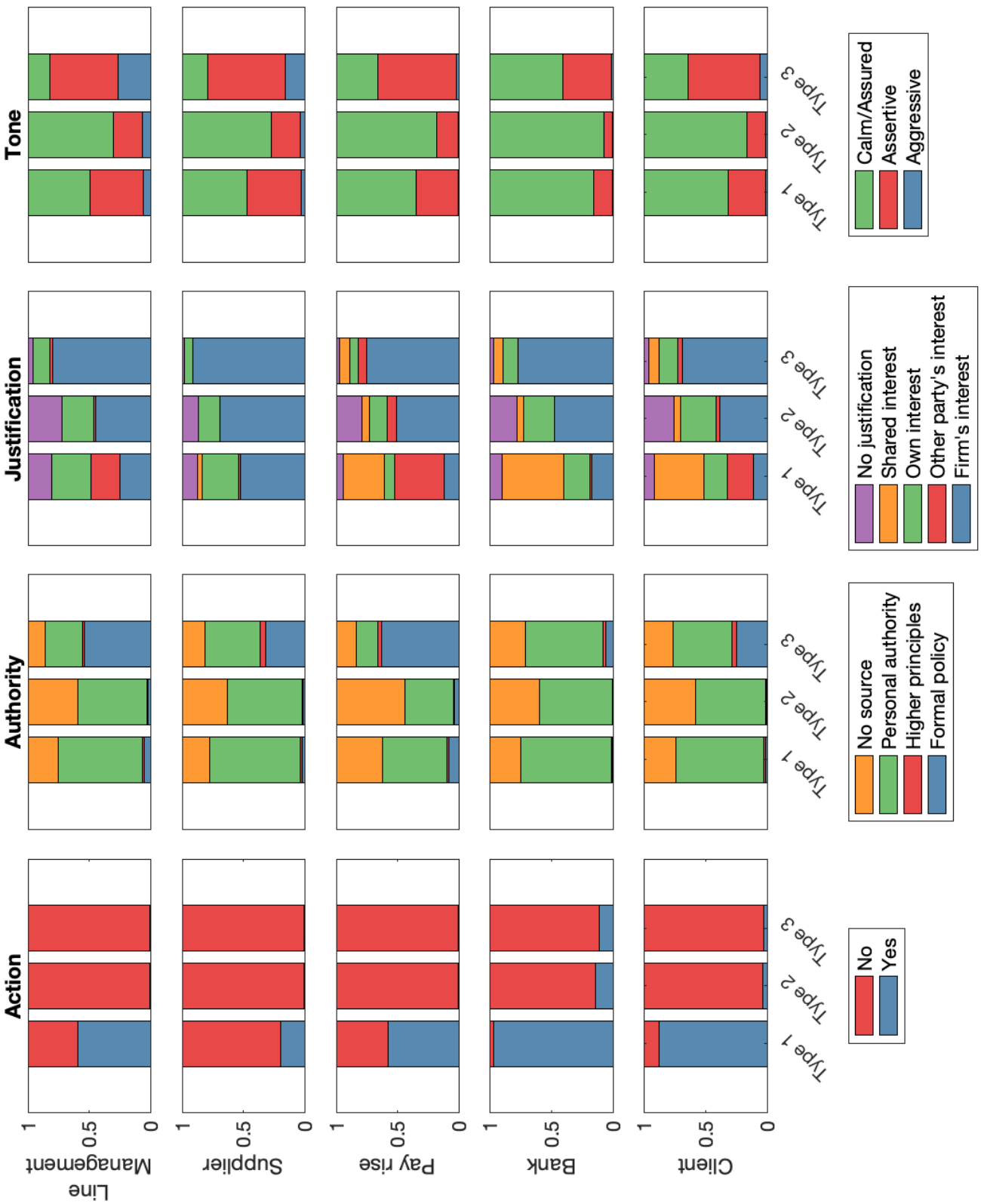


Figure 3.13: 'pure type' management styles amongst Ethiopian young professionals (3-type model)

Figure 3.14: Distribution of types across individuals (3-type model)

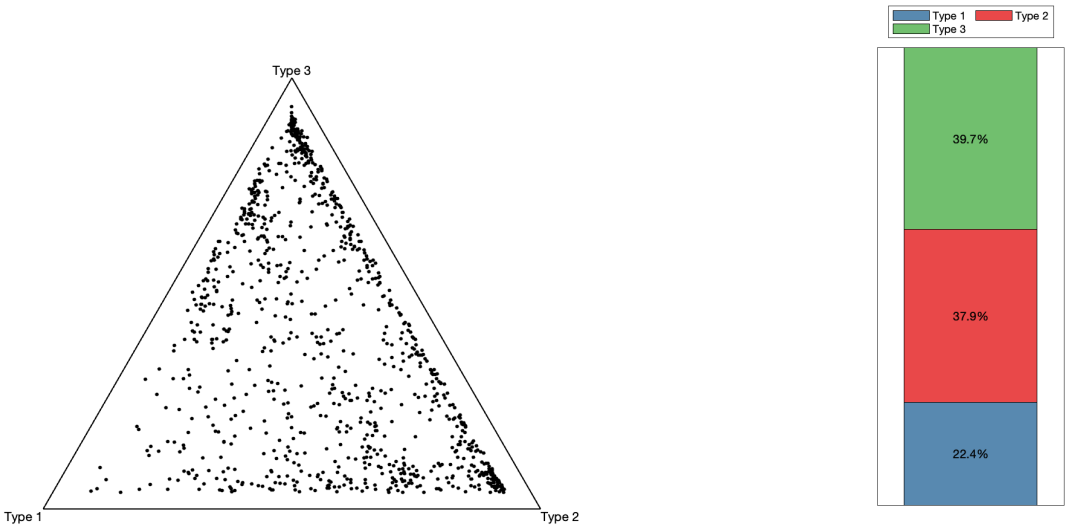


Table 3.19: Agreement in rankings between the respondents' and the HR consultant's rankings. (3-type model)

	Prediction Senior manager and senior manager	Own action and Senior Manager	Internal Agreement
Overall	.511	.415	.793
Type 1	.502	.414	.811
Type 2	.457	.376	.812
Type 3	.565	.451	.768

Notes This table displays Kendall's Tau values comparing the respondents' rankings and the expert's rankings. The column "Prediction of Senior Manager versus Actual Senior Manager" compares the respondents' predictions of the senior manager's actions with the actual senior manager's actions. The column "Own Action versus the Senior Manager" compares the senior manager's actions with what the respondent would do themselves. The column "Internal Agreement" displays Kendall's Tau for the respondents' perception of their own actions versus what they expect a senior manager would do. Kendall's Tau is calculated as $\tau = \frac{\sum \text{agreements} - \sum \text{disagreements}}{\sum \text{pairwise comparisons}}$.

Table 3.20: Characteristics and Types: Summary Statistics (3-type model)

Type	Gender [1=male]	Wage employment indicator	Self-Employment indicator	Above Median Reservation Wage	Above Median Reservation Profit	Average Duration Response
Type 1	.772	.676	.154	.34	.31	38.4
Type 2	.784	.671	.143	.266	.312	36.7
Type 3	.83	.701	.185	.436	.425	53.1
p-value	.042	.476	.152	<0.001	<0.001	< 0.001

Notes This table describes the average characteristics of individuals of each type. Specific, this includes their gender (1 indicating male, 0 female), a dummy for their wage- and self-employment status, and the probability they have an above-median reservation wage and profit based on data collected before respondents' attended the studio. We also include a dummy for whether or not the individual was treated in the management placement experiment, and the average duration of the responses of the candidate across the vignettes. Note that the median splits on reservation wage and profit do not yield a 50/50 split due to bunching in the underlying data at ETB10.000. These numbers are calculated by assigning each individual to the pure type for which they have the highest estimated $\hat{\boldsymbol{\theta}}_i$. Then, a conditional average is taken for each pure type. To test whether intentions differ across types, we compute the Mahalanobis (Wald) statistic and compare it to a χ^2_3 distribution. This relies on the multivariate normal approximation to the posterior described in Section 4.1 of Gelman et al. (2013). The resulting p-value appears in the final row. For non-binary outcomes, we first create a binary split at the median to split the sample.

Table 3.21: Labour market experience and Types: Summary Statistics (3-type model)

Type	Employment	Permanent employment	Unemployed	Management Position	Number of transitions
	Years	*Years*	*Years*	*Share*	*Count*
Type 1	5.292	3.629	.466	.121	.879
Type 2	5.178	3.545	.573	.114	.955
Type 3	5.484	3.864	.363	.185	.688
p-value	.053	.015	.002	.002	.021

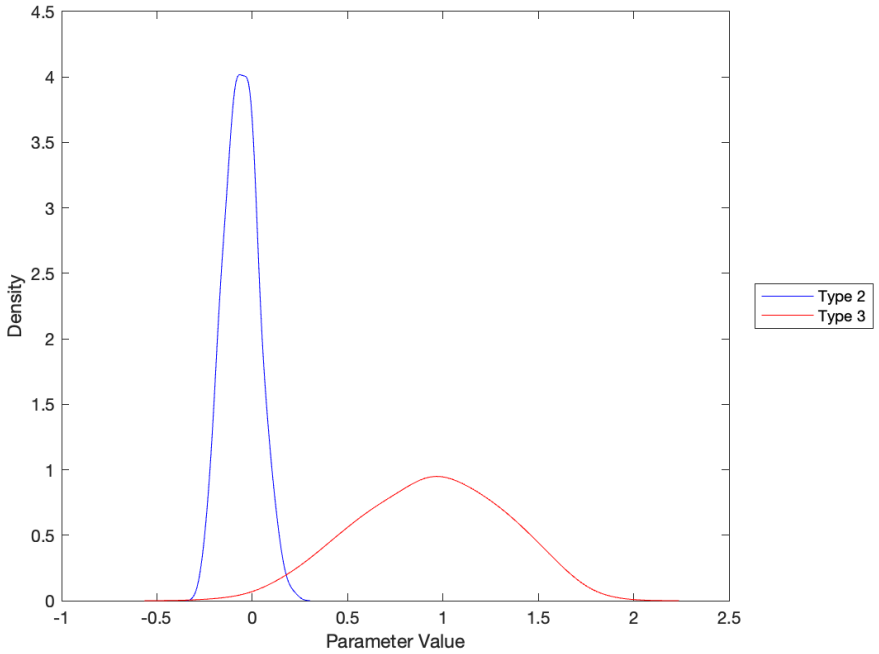
Notes This table describes the labour market experience over the past 6 years of individuals of each type. Specific, this includes the number of years they have been in employment including both self- and wage-employment, the number of years they have been in permanent employment, the number of years they have been unemployed, whether they were in a management position before participating in the studio experiment, and finally the number of labour market transitions they have gone through. These numbers are calculated by assigning each individual to the pure type for which they have the highest estimated $\hat{\boldsymbol{\theta}}_i$. Then, a conditional average is taken for each pure type. To test whether intentions differ across types, we compute the Mahalanobis (Wald) statistic and compare it to a χ^2_3 distribution. This relies on the multivariate normal approximation to the posterior described in Section 4.1 of Gelman et al. (2013). The resulting p-value appears in the final row. For non-binary outcomes, we first create a binary split at the median to split the sample.

Table 3.22: The causal effect of managerial experience on management style by parents education (3-type model)

	Full sample		Low parental education		High parental education	
	(1)	(2)	(3)	(4)	(5)	(6)
Type 1 (%)	21.9	1.4	22.9	1.1	21.1	1.6
		[-0.7, 3.6]		[-2.0, 4.0]		[-1.2, 4.4]
Type 2 (%)	40.4	-4.6	43.7	-7.0	37.3	-2.4
		[-6.8, -1.9]		[-10.7, -3.4]		[-5.8, 0.8]
Type 3 (%)	37.7	3.2	33.4	5.8	41.6	0.8
		[0.6, 5.9]		[2.1, 9.4]		[-2.4, 4.0]
N	479	500	229	239	250	261

Notes This table reports the treatment effect of the management experience experiment on the managerial traits of individuals. The treatment effect is calculated based on the distribution of the difference in the average value of θ for treated and untreated individuals. Columns (1), (3) and (5) report the average estimated value of $\boldsymbol{\theta}_i$ for individuals that were not treated in the management experience experiment for respectively all individuals, individuals whose parents did not finish primary school and for individuals for whom at least one parent did. Columns (2), (4) and (6) report the treatment effect of the management experience experiment on their managerial traits for these three groups respectively. In columns (2), (4) and (6) both the average treatment effect and the 95% credible interval, in square brackets, are reported.

Figure 3.15: The distribution of the preferences for entry-level managers (3-type model)



Notes This figure shows the distribution of the estimates for $\hat{\boldsymbol{\beta}}_{fs}$ across all firms in terms of demand for entry-level managers. In this figure, the Type 1 is omitted and the estimates for $\hat{\beta}_{\text{Type 2}}$, $\hat{\beta}_{\text{Type 3}}$ are plotted

Table 3.23: Actors' gender and management styles (3-type model)

	Panel A: Effect on estimated types		
	Theta 1	Theta 2	Theta 3
First actor female	-0.012	-0.009	0.020
Constant	0.232***	0.385***	0.384***
Bayesian Credible Interval	[-.028 .017]	[-.042 .007]	[0.000 .046]
N	982	982	982

	Panel B: Effect on attributes			
	Justification		Authority	
	Rely on firm's interest	Rely on own interest	Rely on formal authority	Rely on seniority
First actor female	0.032*	-0.028**	0.021	-0.026
	(0.019)	(0.014)	(0.015)	(0.020)
Constant	0.554***	0.178***	0.154***	0.537***
	(0.013)	(0.010)	(0.010)	(0.014)
Enumerator FE	Yes	Yes	Yes	Yes
Vignette FE	Yes	Yes	Yes	Yes
Mean dep. var	0.552	0.168	0.163	0.511
N	6887	6887	6887	6887

	Panel C: Effect on managers' assessments			
	Ranking Data		Normalised likert score	
	Manager	Entrepreneur	Manager	Entrepreneur
First actor female	0.052**	0.060***	0.083**	0.093***
	[0.010 0.093]	[0.017 0.102]	(0.033)	(0.033)
Constant	0.474***	0.470***	-0.045*	-0.050**
	[0.453 0.493]	[0.449 0.492]	(0.025)	(0.025)
Vignette FE			Yes	Yes
N			6874	6869

Notes: This figures displays the causal link between the first actor a respondent sees and their subsequent responses (for the second to fifth vignette). Panel A shows that this indeed manifests as a reliance on a more authoritative management style (captured approximately by type 3), reporting 95% Bayesian Credible Intervals in square brackets. Panel B shows that respondents act more like an authoritative type - they agree less, rely on formal policy and authority - after starting with a female actor. These attributes are selected to illustrate the shift from a coercive to an authoritative management style; Appendix 3.7.L reports a sequence of multinomial logit regressions on the full set of attributes from which these were selected. The third panel shows the effect of (a) the first actor being female and (2-3) the results from an Acharya-style mediation analysis for using first the estimates for the type parameter from Panel B) and then the attributes of the responses from Panel A. Statistical significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

3.7.F.4 Considering five types

Finally, we consider setting $K = 5$. There are several indications that five types are excessive in this context. First, the five-type model is bimodal – in the sense that we can find two distinct kinds of classification that have approximately equal log-likelihoods (*i.e.* bimodality in the substantive description of types – not merely bimodality due to 'label-switching'). Second, one of these two alternative classifications is descriptively very similar to the results from the four-type model: in essence, the five types could be described as 'authoritative', 'affiliative', 'coercive', 'timid' and 'timid-coercive' – and, of these, the 'timid' and 'timid-coercive' behave in extremely similar ways. Third, neither of these alternative classifications seems to add any additional nuance or insight to the conclusions reached from the four-type model.

3.7.G Inclusion in the studio

Table 3.24: Logit Regression Results on Studio Attendance and Management Experience Treatment

	(1) Baseline sample	(2) 4-year follow-up sample	(3) Studio sample
	Attended Studio	Attended Studio	Treated
Treated	0.0208	0.0159	
	(0.0239)	(0.0241)	
Gender (1=Female)	-0.138***	-0.121***	0.0783*
	(0.0275)	(0.0283)	(0.0413)
Age at baseline	-0.00425	-0.000840	-0.00149
	(0.00435)	(0.00456)	(0.00589)
BA Degree	-0.109*	-0.0175	0.0142
	(0.0591)	(0.0545)	(0.0706)
Either parent finished primary school	-0.0135	-0.000208	-0.0128
	(0.0243)	(0.0252)	(0.0327)
Either parent owned a business	-0.0311	-0.0123	-0.00308
	(0.0247)	(0.0252)	(0.0333)
In wage employment		0.00745	0.0664
		(0.0318)	(0.0420)
Manages others while in wage emp.		0.0635	-0.0261
		(0.0420)	(0.0508)
In self-employment		-0.0127	0.0609
		(0.0395)	(0.0516)
Above median reservation wage		-0.0726**	0.00834
		(0.0294)	(0.0390)
Above median reservation profit		0.0470	0.0346
		(0.0299)	(0.0383)
Observations	1637	1429	993

Notes: This table reports the average marginal effects from a logit regression on attending the studio conditional of being part of the previous management experience experiment in column (1) and conditional on being part of the four-year follow-up in column (2). Finally, column (3) reports the average marginal effects from a logit regression on being treated conditional on attending the studio in column (2). As independent variable a set of variables related to socioeconomic background of studio participants, and labour market outcomes before attending the studio are included. Robust standard errors in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.25: Inverse Probability Weighted Mean of θ

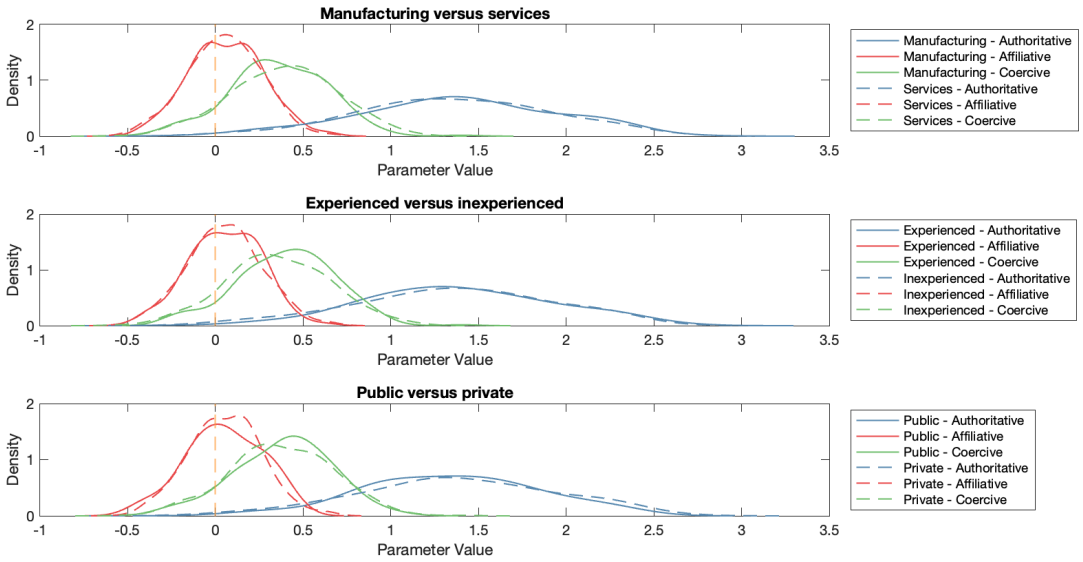
	Unweighted	Weighted
Timid	.252	.252
Authoritative	.278	.276
Affiliative	.190	.190
Coercive	.280	.281

This table reports the unweighted mean of θ, and the mean of θ after an inverse probability weighting. This result suggests that the individuals included in the studio sample are representative of the full sample of individuals that were invited to participate in the studio experiment. The weights are calculated based on the probability of attending the studio as a function of a set of baseline characteristics from before the management experience experiment, including wage and self-employment status, gender, age, a dummy whether the individual has a bachelor's degree, a dummy for whether the individual has any intentions of starting a business, parents' education (a dummy equal to one if either parent finished primary school), and parents' experience running a business (a dummy equal to one if either parent ran a business).

3.7.H Variations on the Plackett-Luce model

3.7.H.1 The distribution of β_f conditional on covariates

Figure 3.16: The distribution of $\boldsymbol{\beta_f}$ conditional on covariates



Notes This figure shows the distribution of the estimates for $\hat{\boldsymbol{\beta}}_{fs}$ for a number of subsamples of the data. The top pane splits the sample by whether the firm operates in a manufacturing or services sector, the second pane splits the sample by the median number of years of experience of the manager, and the third pane splits the sample by whether the firm is public or private.

3.7.H.2 Allowing firm preferences to vary by vignette

$$U_{fiv} = \boldsymbol{\beta}_{fv}\boldsymbol{\theta}_i + \gamma_i + \varepsilon_{fiv}, \tag{3.14}$$

The priors for the Dirichlet model and γ_i are the same as in our main specification. For the parameters $\boldsymbol{\beta}_{fv}$ we specify the following priors:

$$\boldsymbol{\beta}_{fv} \sim \text{MVN}\left(\boldsymbol{\beta}_v, \Sigma_v\right), \quad \text{with } \Sigma_v = \text{diag}(\boldsymbol{\sigma}_v)\,\mathbf{I}\,\text{diag}(\boldsymbol{\sigma}_v),$$

$$\boldsymbol{\beta}_v \sim \mathcal{N}\left(\mathbf{0},\ (K-1)\,\mathbf{I}\right),$$

$$\sigma_{v,j} \sim \text{Uniform}(0, 10) \quad \text{for } j = 1, \ldots, K-1.$$

Table 3.26: Heterogeneity demand for types by vignette

	Employee Absence	Supplier	Pay Rise	Bank	Client
Authoritative	2.374	1.803	1.892	1.304	1.656
	[1.552, 3.381]	[1.094, 2.671]	[1.235, 2.666]	[0.690, 1.994]	[0.976, 2.478]
Affiliative	0.254	0.035	-0.345	-0.038	0.008
	[-0.533, 1.037]	[-0.809, 0.893]	[-1.171, 0.472]	[-0.810, 0.762]	[-0.767, 0.809]
Coercive	0.767	-0.141	0.715	0.059	0.737
	[0.076, 1.504]	[-0.795, 0.497]	[0.020, 1.492]	[-0.498, 0.657]	[0.100, 1.450]

3.7.H.3 Allowing firm preferences to vary by candidate gender

$$U_{fiv} = \boldsymbol{\alpha}_f \texttt{I[female]}_i + \boldsymbol{\beta}_f \boldsymbol{\theta}_i + \boldsymbol{\delta} \cdot \texttt{I[female]}_i \cdot \boldsymbol{\theta}_i + \gamma_i + \varepsilon_{fiv}, \tag{3.15}$$

The priors for the parameters α_f, β_f and γ_i are the same as in our main specification, as well as all priors in the Dirichlet model. For the new parameters, i.e. $\boldsymbol{\delta}$, we specify the following priors:

$$\boldsymbol{\delta} \sim \mathcal{N}(0, 3);$$

Table 3.27: Heterogeneity by the gender of the candidate

Parameter	Estimate	90% Credible Interval
β_1	1.253	[0.925, 1.595]
β_2	0.205	[-0.254, 0.653]
β_3	0.286	[-0.039, 0.628]
α_1	-0.092	[-0.599, 0.410]
δ_1	0.692	[-0.055, 1.449]
δ_2	-0.457	[-1.396, 0.495]
δ_3	0.411	[-0.293, 1.141]

3.7.H.4 Overcontrolling for the duration of the response

We also re-estimate our main model controlling for the duration of the vignette to assess whether authoritative candidates are preferred simply because they speak for longer. Of course, verbosity may itself reflect the type of behaviour we aim to capture, so this constitutes an over-control. Nonetheless, it serves as a useful robustness check for whether duration mechanically drives the results.

Specifically, we estimate the following equation, allowing the coefficient on duration, $\boldsymbol{\alpha}$, to vary by vignette. We place an independent normal prior on each element of $\boldsymbol{\alpha}$, with mean zero and variance three.

$$U_{fiv} = \boldsymbol{\alpha} \cdot \texttt{duration}_{iv} + \boldsymbol{\beta}_f \boldsymbol{\theta}_i + \gamma_i + \varepsilon_{fiv}$$

Table 3.28: Demand (over)controlling for duration

Parameter	Estimate	90% Credible Interval
$\beta_{authoritative}$	0.875	[0.567, 1.199]
$\beta_{affiliative}$	-0.012	[-0.394, 0.355]
$\beta_{coercive}$	0.231	[-0.059, 0.535]
α_1	0.010	[0.006, 0.013]
α_2	0.006	[0.002, 0.009]
α_3	0.009	[0.006, 0.012]
α_4	0.007	[0.003, 0.010]
α_5	0.008	[0.004, 0.012]

3.7.I Describing heterogeneity in firm preferences

To assess how heterogeneous firms are in their revealed preferences for aspiring manager, we first assess the probability that two managers agree with each other, both for the full sample and by individual vignette. Table 3.29 shows that the probability of agreement between two managers varies between 56% and 61% across the vignettes. To quantify the level of agreement between firms, we implement a random-effects rank-ordered logit model without covariates; we specify the latent utility from firm f assessing candidate i

as:

$$y^*_{fivs} = \mu_{ivs} + \varepsilon_{fivs},$$

where we specify $\mu_{ivs} \sim \mathcal{N}(0, \sigma^2_\mu)$ and we assume that ε_{fivs} has a Type 1 Extreme Value distribution.[34] We assume that μ_{ivs} is independent across vignettes and types of employment (so, in effect, we estimate this model separately for each vignette and each type of employment).[35] This model reveals much about firm preferences and the notion of 'management as technology' because the variance σ^2_μ captures the strength of relative agreement among firms. In one limiting case, firm preferences are completely idiosyncratic: $\sigma^2_\mu = 0$. In another limiting case, preferences are common across all firms: $\sigma^2_\mu \to \infty$. An important intermediate case is where the proportion of variation due to common preferences equals the proportion due to idiosyncratic preferences: $\sigma^2_\mu = \pi^2/6$. In this way, estimation of σ^2_μ provides a specific figure — to our understanding, the first formal quantification in the literature — for heterogeneity in firm preferences over management traits.[36]

We find that, in line with the probability of agreement between managers, the share of the variation due to idiosyncratic preferences by far outweighs that share of the variation explained by common preferences. Focusing on the agreement on the ranking as a manager, we estimate that around 20% (for the line management vignette) to 35% (for the supplier vignette) of the total variation in the latent utility is explained by the common component μ.

[34] We again estimate this using Hamiltonian Monte Carlo using the Stan language.
[35] We refer to self-employment and wage-employment as two distinct types of employment throughout.
[36] This approach is broadly analogous to the identification strategy of Bertrand and Schoar (2003), which involves comparing R^2 values when adding manager fixed effects; this is conceptually similar to the role played by μ_i.

Table 3.29: Dyadic agreement of managers across vignettes

	Vignette					
	Line Management	**Supplier**	**Pay rise**	**Bank**	**Client**	**Total**
Ranking as manager						
Probability of agreement	56.7	60.9	58.0	57.7	59.3	58.5
Variance μ	0.26	0.78	0.41	0.48	0.60	0.58
Ranking as entrepreneur						
Probability of agreement	57.9	57.9	56.3	60.2	60.5	58.5
Variance μ	0.43	0.53	0.25	0.70	0.74	0.52

Notes This table depicts the probability of agreement and the variance of μ. The first is the probability that two managers assessing the same pair of candidates agree on their ranking as an entrepreneur and as a manager, split by the two types of employment and across vignettes. The variance of μ provides a measure of the homogeneity of preferences across managers.

To decompose the heterogeneity in preferences for management traits across firms, we first analyse whether the dyadic agreement between firms can be predicted by observable characteristics. To do so, we implement median splits of the sample based on a number of observable characteristics. We then create three indicator variables: $\texttt{low}_{ij}$, equal to one if both firms are below median for that characteristic, $\texttt{different}_{ij}$ equal to one if one firm is below, and one firm above median, and $\texttt{high}_{ij}$ if both firms are above the median.

We then analyse the resulting data using the following regression model for pairs of firms $i \neq j$ with standard errors clustered at the pair level:

$$\texttt{agree}_{ij} = \alpha + \beta_1 \texttt{different}_{ij} + \beta_2 \texttt{high}_{ij} + \varepsilon_{ij} \tag{3.16}$$

Table 3.30 reports the results from this regression. These results further stress that agreement between the HR managers is relatively low, and show the relationship with observable characteristics is limited. We see effectively no differences in rates of agreement by management score as calculated based on MOPS-type questions, number of employees, share of employees in a management position at the firm and share of non-payroll

employees (in respectively columns 1, 2, 4 and 5). We only observe a clear increase in two characteristics: the number of competitors and managers' self-reported trust. Firms in more competitive environments appear to be more likely to agree (column 1), where the effect seems to be additive although only significant when both firms have a high number of competitors. The most clear relationships seems to be for high self-reported trust, an indicator for whether the manager reports to be generally trusting. We find that the rate of agreement drops sharply when either manager reports to be trusting (column 6). This suggests that non-trusting managers pick up on some dimension of the responses, and trusting managers ignore this dimension without coordinating on some other observable instead. This is a sharp difference, with either manager being trusting decreasing the probability of agreement from 62.8% to around 55% in a binary choice.

Table 3.30: Agreement between HR managers by firm and manager characteristics

Split by:	(1) Management Score	(2) Number of employees	(3) Number of competitors	(4) Share of employees in management	(5) Share of non-payroll employees	(6) Manager self-reported trust
One high, one low	0.00885	0.0252	0.0331	-0.00837	0.000646	-0.0712
	(0.0235)	(0.0245)	(0.0230)	(0.0261)	(0.0216)	(0.0223)
Both high	0.00226	0.0392	0.0687*	0.0218	-0.0166	-0.0831**
	(0.0271)	(0.0289)	(0.0326)	(0.0287)	(0.0317)	(0.0299)
Constant	0.580***	0.563***	0.558***	0.584***	0.587***	0.628 ***
	(0.0182)	(0.0202)	(0.0190)	(0.0217)	(0.0141)	(0.0163)
N	3948	3948	3849	3860	3948	3553

Notes This table reports the parameter estimates of regression 3.16. The column labels indicate the variables based on which the binary split of the sample for each regression is made. The dependent variable is the share of agreement between the two HR managers assessing the same pair of candidates. The variables on which the splits are made are the overall management score, the number of employees, the number of competitors, the share of employees in management positions, the share of non-payroll employees, and the self-reported trust of the manager. The number of observations in column (6) is lower due to some observations missing for this variable. The standard errors are in parentheses and the p-values are reported as: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.7.J Worker preferences over managers

This paper shows that firm managers strongly favor the authoritative managerial type. A natural follow-up is whether employees – who would work under such managers – hold

similar preferences. To explore this, we carried out a supplementary data-collection exercise with workers supervised by entry-level managers, presenting them with the vignette recordings and gathering their evaluations through a series of closed- and open-ended questions.

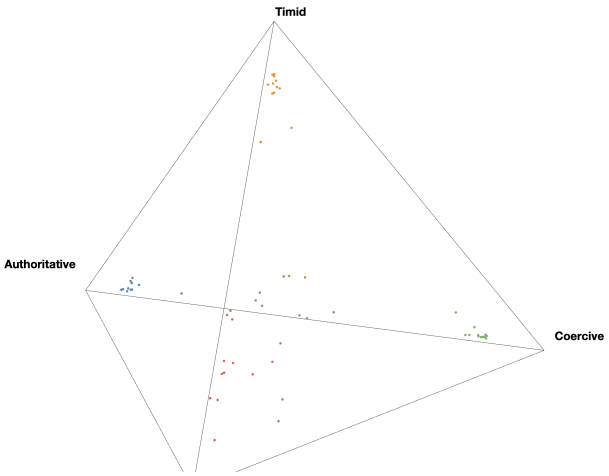
Specifically, in May 2025 we revisited 17 firms from our sample.[37] At each firm, we asked the on-site manager to share the phone numbers of up to five of their employees in their main production process, selecting those who are managed by the lowest level of managers in the firm. We then contacted these workers and asked them to participate in a focus group in central Addis Ababa.

We ran these focus groups with groups of four to seven participants. We focused on the scenarios related to managing employees, so the line-management and pay-rise scenarios. We designed the exercise to be model-informed – after estimating the model – showing, for each of the two vignettes, recordings of four 'archetypal' managers and one 'average' manager. The "archetypal" managers are those most closely resembling each of the pure types in our sample. For example, for the authoritative type we shared the contact details of the most authoritative managers in our sample based on $\boldsymbol{\theta_i}$, specifically from the 95th percentile of the distribution. We then asked an enumerator to contact them going from most to least authoritative within this subsample to ask for consent to show their videos to workers. We repeated this for each pure type. In practice, we always selected among the 20 managers closest to each pure type. For the "average" manager, we created this list by first calculating the entropy of classifying each individual based on $\boldsymbol{\theta_i}$ and contacted the managers with the highest classification uncertainty. Figure 3.17 shows the distribution of $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$ for individuals included in the simplex, showing they are drawn from either the corners of the simplex – the pure types – and from the centroid of the simplex - the average types. Authoritative, coercive, and timid managers are all close to the corners of the simplex, whereas affiliative managers are further from their respective corner in terms of $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$ as there are few "pure" affiliative managers.

Because workers might evaluate male and female managers differently for reasons

[37] After contacting the IRB at Oxford University to confirm our updated protocol was appropriate and covered by our existing ethics approval.

Figure 3.17: Distribution of $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$ for individuals in focus groups



unrelated to managerial style, we stratified the vignettes by manager gender. Given our modest sample size, each participant viewed recordings featuring exclusively male or exclusively female managers: the first nine focus groups saw male managers and the final three saw female managers. This design mirrors the composition of our sample of young professionals, in which 75 percent of participants are male.

The average worker in our sample was 31.2 years old, had completed 12.8 years of schooling, had spent four years with the current firm and accumulated 7.5 years of labour-market experience (Table 3.31). The sample was gender balanced (47.0 percent male). Most respondents worked in manufacturing (51.5 percent) or services (42.4 percent), with the remainder spread across smaller sectors. Their roles were concentrated in sales, administrative, and other customer-facing functions rather than in production.[38] Additional summary statistics appear in Table 3.32.

Table 3.31: Summary Statistics – Continuous Variables

	mean	p50	sd	min	max
Age	31.2	30	9.7	19	57
Years of education	12.8	12	2.5	8	18
Experience with firm (years)	4.0	2	4.9	1	26
Labour market experience (years)	7.5	6	7.1	1	31
Observations	66				

[38] Weekday scheduling initially excluded production workers, who were indispensable to ongoing operations. After the first recruitment wave, most sessions were therefore moved to Saturdays, which are typically half-days in Ethiopia.

Table 3.32: Distribution of categorical demographic variables

Variable	Category	Count	Percent
Gender	1 = Male	31	47.0
Gender	2 = Female	35	53.0
Position	1. Intern / Trainee / Apprentic	1	1.5
Position	2. Administrative / Clerical Assistant	9	13.6
Position	3. Customer-facing Service / Teller	14	21.2
Position	4. Sales / Marketing Assistant	29	43.9
Position	5. Technical / IT Support Junior	7	10.6
Position	6. Production / Operator / Field Assistant	5	7.6
Position	7. Junior Professional (Associate / Analyst)	1	1.5
Sector	1 = Manufacturing	34	51.5
Sector	2 = Services	28	42.4
Sector	3 = Finance	2	3.0
Sector	4 = ICT	1	1.5
Sector	5 = Public/NGO	1	1.5
Education	Primary school	2	3.0
Education	High school	22	33.3
Education	Vocational school	7	10.6
Education	Diploma	10	15.2
Education	BA	21	31.8
Education	MSc	4	6.1

Turning to the assessments. In the first part of this exercise, we asked the respondents to, separately for the two scenarios, watch five videos of young professionals responding to the scenario. After showing the respondents each video, we asked them individually to (a) list up to three words describing that person and (b) five Likert-style questions on a 5-point scale from Strongly Disagree to Strongly Agree (listed below):[39]

1. **Productivity** This person would raise the productivity of teams in my firm.
2. **Targets** This person would help the firm achieve its targets.
3. **Motivating** I would be happy to work hard for this person.
4. **Consistent** This manager would treat different employees consistently.
5. **Typical** This type of response is typical for a manager in my firm.

After watching all five videos for one vignette, we asked the respondents to rank these five videos; to remind respondents which managers they had seen, we showed them photos of the five managers at this point. Enumerators reported that respondents were generally able to recall the managers based on these screenshots. We specifically asked the following question:

> Now, we would like you to rank these five responses from the most effective manager to the least effective manager; by effective we mean someone who would contribute to the long-term success of the firm he works for. We will show you photos of each of the aspiring managers you just saw and ask you to rank them.

We estimate two models using Stata to analyze responses: a rank-ordered logit model for rankings and an ordered logit model for Likert-scale ratings.

```
rologit rank i.type, group(respondentVignetteUniqueID) reverse
```

where `rank` is the rank from the respondent from one to five, i.type a dummy for being each "pure" type or the average type, and respondentVignetteUniqueID groups data at

[39] The label in bold text is reported in all regression tables as a shorthand.

the respondent-vignette level. We report the estimates in ordered logit coefficient space.

For the Likert data:

```
ologit Y i.type , cluster(respondentUniqueID)
```

where Y_i is the score from the respondent, i.type a dummy for being each "pure" type or the average type, and respondentUniqueID a respondent ID. We currently report results in logit space.

Table 3.33 provides the main results from this exercise, where the Timid type is omitted. Respondents show a clear preference for authoritative managers among the five types. This pattern is reflected in both the ranking data in column 1, as well as for each individual Likert question. When we look at the remaining types, we see both the Coercive and Average type are perceived as better than the Timid type, and largely for very similar reasons. Interestingly, the Affiliative type is perceived as the worst type.

Table 3.33: Workers' perceptions of management styles

	(1) Ranking	(2) Productivity	(3) Targets	(4) Motivating	(5) Consistent	(6) Typical
Authoritative	0.938***	1.292***	1.468***	1.039***	1.286***	0.720***
	(0.163)	(0.241)	(0.257)	(0.240)	(0.226)	(0.213)
Coercive	0.387**	0.247	0.526**	0.208	0.218	0.277
	(0.156)	(0.243)	(0.229)	(0.198)	(0.240)	(0.217)
Affiliative	-0.269*	-0.801***	-0.595**	-0.649***	-0.652***	-0.714***
	(0.158)	(0.223)	(0.244)	(0.220)	(0.209)	(0.215)
Average	0.429***	0.308	0.383*	0.460**	0.416*	0.308
	(0.157)	(0.196)	(0.225)	(0.194)	(0.213)	(0.201)
Cutoff 1		-2.832***	-3.177***	-2.121***	-2.666***	-2.314***
		(0.232)	(0.272)	(0.174)	(0.197)	(0.202)
Cutoff 2		-0.985***	-0.986***	-0.823***	-0.997***	-0.581***
		(0.178)	(0.200)	(0.162)	(0.184)	(0.170)
Cutoff 3		-0.253	-0.194	-0.154	-0.146	0.143
		(0.194)	(0.200)	(0.161)	(0.174)	(0.154)
Cutoff 4		1.726***	2.052***	1.925***	2.017***	2.520***
		(0.231)	(0.238)	(0.210)	(0.212)	(0.232)
Auth.=Coercive	0.001	0.000	0.001	0.002	0.000	0.068
Auth.=Affiliative	0.000	0.000	0.000	0.000	0.000	0.000
Auth.=Average	0.001	0.000	0.000	0.015	0.000	0.074

Notes: Standard errors clustered by respondent. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Finally, we implement this for the two vignettes separately. We report the results in Tables 3.34 and 3.35 for the line management and pay rise scenario respectively. We find broadly similar trends across the two scenarios, although the timid type appears to perform relatively better in the line management scenario than in the pay rise scenario when compared to the other types.

Table 3.34: Line Management: Worker's perceptions of management styles

	(1) Productivity	(2) Targets	(3) Motivating	(4) Consistent	(5) Typical	(6) Ranking
Authoritative	1.397***	1.753***	0.922***	1.629***	0.626**	0.684***
	(0.334)	(0.363)	(0.328)	(0.334)	(0.275)	(0.231)
Coercive	0.498	0.897***	0.205	0.581	0.373	0.278
	(0.338)	(0.319)	(0.305)	(0.357)	(0.343)	(0.230)
Affiliative	-0.825***	-0.577*	-0.767**	-0.546*	-0.568*	-0.529**
	(0.291)	(0.326)	(0.308)	(0.297)	(0.326)	(0.228)
Average	0.202	0.261	0.240	0.588*	0.266	0.211
	(0.306)	(0.344)	(0.309)	(0.308)	(0.338)	(0.222)
Cutoff 1	-2.753***	-3.195***	-2.320***	-2.742***	-2.450***	
	(0.287)	(0.415)	(0.253)	(0.316)	(0.243)	
Cutoff 2	-0.971***	-0.771***	-0.966***	-0.859***	-0.555**	
	(0.222)	(0.230)	(0.207)	(0.236)	(0.226)	
Cutoff 3	-0.254	-0.0982	-0.359*	-0.109	0.156	
	(0.232)	(0.235)	(0.215)	(0.227)	(0.230)	
Cutoff 4	1.850***	2.216***	1.859***	2.336***	2.577***	
	(0.292)	(0.291)	(0.261)	(0.281)	(0.320)	
Observations	305	305	305	305	305	305

Notes: Using only data from the line management scenario. Standard errors clustered by respondent. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.35: Pay Rise: Worker's perceptions of management styles

	(1) Productivity	(2) Targets	(3) Motivating	(4) Consistent	(5) Typical	(6) Ranking
Authoritative	1.206***	1.205***	1.151***	1.021***	0.821**	1.184***
	(0.312)	(0.319)	(0.281)	(0.298)	(0.324)	(0.232)
Coercive	0.0239	0.179	0.221	-0.0845	0.195	0.491**
	(0.342)	(0.340)	(0.271)	(0.302)	(0.307)	(0.215)
Affiliative	-0.776**	-0.628*	-0.546*	-0.749***	-0.856***	-0.0287
	(0.323)	(0.333)	(0.323)	(0.290)	(0.295)	(0.220)
Average	0.387	0.464	0.655**	0.270	0.342	0.641***
	(0.264)	(0.301)	(0.272)	(0.311)	(0.288)	(0.223)
Cutoff 1	-2.917***	-3.192***	-1.947***	-2.626***	-2.209***	
	(0.339)	(0.358)	(0.268)	(0.276)	(0.302)	
Cutoff 2	-1.000***	-1.209***	-0.692***	-1.120***	-0.607**	
	(0.260)	(0.291)	(0.229)	(0.244)	(0.246)	
Cutoff 3	-0.250	-0.295	0.0339	-0.175	0.132	
	(0.263)	(0.269)	(0.214)	(0.236)	(0.225)	
Cutoff 4	1.626***	1.919***	1.991***	1.764***	2.475***	
	(0.262)	(0.300)	(0.251)	(0.262)	(0.290)	
Observations	329	329	329	329	329	329

Notes: Using only data for the pay rise scenario. Standard errors clustered by respondent. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.7.K Summary statistics

Table 3.36: Descriptive statistics of young professionals

Categorical Variables				
Gender (% Male)	78%			
University degree	78%			
Self-employed	17%			
Wage-employed	68%			
Neither parent completed university	88%			
Neither parent completed primary school	48%			
Conditional on Wage Employment				
In a professional position	98%			
In a managerial position	18%			
Continuous Variables				
	Median	Std. Dev.	5th Perc.	95th Perc.
Age (years)	31	2.8	27	36
Wage (ETB/Month)	6000	5979	2285	18000
Profit (ETB/Month)	6000	16099	0	42000
N	982			

Notes This table provides summary statistics of the respondents included in the sample. The wage and profit in Ethiopian Birr (ETB) are calculated conditional on being in wage- and self-employment, respectively. The percentages of individuals in professional and managerial positions are conditional on being in wage employment.

Table 3.37: Descriptive statistics of firm managers

Categorical Variables				
BA Degree	79.7%			
MSc Degree	24.0%			
Formal management education	72.4%			
Department in firm				
Finance	12.6%			
HR	40.5%			
Administration	34.6%			
Other	12.3%			

Continuous Variables				
	Median	Std. Dev.	5th Perc.	95th Perc.
Years of experience at this firm	8	8.02	1	27
Years of labour market experience	16	14.56	4	46
N	587			

Notes This table provides summary statistics of the HR managers included in the sample.

Table 3.38: Descriptive statistics of firms

Sector

	Count	Share
Construction	51	8.69%
Education	27	4.60%
Healthcare	30	5.11%
Manufacturing	173	29.47%
Services	167	28.45%
Trade	86	14.65%
Other	53	9.03%

Continuous Variables

	Median	Std. Dev.	5th Perc.	95th Perc.
Number of payroll employees	58	1771.478	6	700
Share of female payroll employees	40.0%	19.2%	11.4%	73.1%
N	587			

Notes This table provides summary statistics of the firms included in the sample.

3.7.L Robustness gender multinomial logit results

In this section, we report the multinomial logit results of regressions of the gender of the actor seen in the first vignette on the attribute exhibited by the respondent in the subsection vignettes.

Table 3.39: Female first actor and actions in subsequent vignettes

	Agree
	b/se
First actor female	-0.120
	(0.097)
Constant	-1.665***
	(0.070)
Observations	6887

Table 3.40: Female first actor and authority in subsequent vignettes

	Source of authority
	b/se
A formal policy	
First actor female	0.179
	(0.112)
Constant	-1.246***
	(0.083)
Higher principles	
First actor female	0.417
	(0.312)
Constant	-3.821***
	(0.242)
Personal relationship	
First actor female	0.678
	(0.429)
Constant	-4.626***
	(0.340)
No source of authority	
First actor female	0.026
	(0.093)
Constant	-0.605***
	(0.067)
Observations	6887

The omitted category is reference to seniority

Table 3.41: Female first actor and justification in subsequent vignettes

	Justification
	b/se
Other party's interest	
First actor female	-0.123
	(0.164)
Constant	-2.428***
	(0.118)
Respondent's best interest	
First actor female	-0.225*
	(0.107)
Constant	-1.134***
	(0.075)
Shared interest	
First actor female	-0.150
	(0.116)
Constant	-1.630***
	(0.082)
No justification h	
First actor female	0.004
	(0.128)
Constant	-1.604***
	(0.096)
Observations	6887

The omitted category is reference to protecting/helping the company.

Table 3.42: Female first actor and tone in subsequent vignettes

	Tone
	b/se
Aggressive	
First actor female	-0.079
	(0.168)
Constant	-2.262***
	(0.119)
Assertive	
First actor female	-0.014
	(0.089)
Constant	-0.401***
	(0.063)
Observations	6887

The omitted category is reference to a calm/assured tone

Social Image, Organisational Values and Inclusion: Evidence from a Field Experiment

Girum Abebe
Siân Brooke
Tom Gole
Simon Quinn
Tom Schwantje

Abstract

We conduct a novel field experiment in Ethiopia to assess institutional barriers to inclusive decision-making. Specifically, we examine how social image concerns and organisational messaging on equal opportunity influence decision-making in a business plan competition. We find that male and female candidates are equally likely to win across all treatment arms, showing no effect on gender equity. However, both interventions increase agreement among judges assessing the same candidates and align individual judges' decisions with experts' preferences. This suggests that the treatments enhance the quality of decision-making among individual judges. Finally, we use machine-learning methods to further explore how the treatments improve the quality of judges' decisions.

4.1 Introduction

In recent years, much public debate has focused on the inclusivity – or lack thereof – of formal institutions. Corporations, universities, and governments are challenged to consider how they can be more diverse and inclusive, particularly concerning gender and the representation of minority groups. Existing research has focused on the attitudes of individual decision-makers (Card, DellaVigna, Funk, and Iriberri, 2019; Exley and Nielsen, 2024), or on the use of tests and algorithmic decision-making (Hoffman, Kahn, and Li, 2018; Li et al., 2020). However, there is a dearth of work in understanding how different institutional structures might hinder or promote inclusivity and material benefits to under-represented groups.

We ran a novel field experiment designed to test institutional barriers to inclusive decision-making. Specifically, we develop this experiment based on a business plan competition involving young professionals in Ethiopia. This competition is assessed by senior human resources managers in large Ethiopian firms, and we vary the conditions under which these managers assess the submissions. In this experiment, we focus on the interaction of two common institutional features: the effect of communicating organisational values (specifically relating to equal opportunity for male and female applicants), and the social image concerns that arise when decisions need to be justified to peers. Our research addresses three key questions. First, *how are evaluators affected by institutional messages promoting equal opportunity?* To answer this, we randomly assign some judges to an 'organisational values' treatment – in which they are exposed to an initial message emphasising the importance of promoting access to capital for women-owned microenterprises. Second, *how do social image concerns influence decision-making?* To answer this question, we randomly assign some judges to a 'social image' treatment, in which they are told in advance that they will be required to discuss and justify their decisions with peer judges. And, third, *how do these factors interact to shape inclusive institutional decision-making?* To answer this question, we implement a third treatment arm that combines both the 'organisational values' prompt and the 'social image' treatment.

We find no impact of any of these three treatments on the average probability of female candidates winning. This result may not be surprising, given that male and female candidates in the control group were equally likely to win. Instead, we find that all three treatments significantly increase the probability that judges agree with each other, and that they agree with the choices of an outside expert. We interpret this agreement – and, in particular, the agreement with the outside expert – as evidence that the treatments increase the quality of assessors' decisions. In particular, we find that the effect on agreement is driven by treated judges agreeing with the experts in those cases where the experts' opinion is strongly expressed; further, we find that treated judges are significantly more likely to care about the substance of candidates' presentations, rather than candidates' age and appearance.

To motivate the experiment and our results, we develop a simple theoretical model in which evaluators choose the identity through which they make decisions, in the spirit of Akerlof and Kranton (2000). In this model, evaluators decide how to weigh organisational objectives against their personal preferences, based both on an intrinsic preference for contributing to organisational values and on social image concerns. This model predicts both treatments could increase the performance of female candidates and agreement amongst evaluators, although through different mechanisms.

Our results contribute to several related bodies of literature. First, we contribute to empirical behavioural literature on the role of social pressure. DellaVigna, List, and Malmendier (2012), Dellavigna, List, Malmendier, and Rao (2016) and Gerber, Green, and Larimer (2008) use randomized field experiments to estimate the importance of social pressure and social image to encourage pro-social behaviour. Ai, Chen, Chen, Mei, and Phillips (2016) and Charness and Holder (2019) study how individual behaviour differs when part of a team versus as an individual. Similarly, several recent experiments highlight the role of social comparisons upon employee performance (Bandiera, Barankay, and Rasul, 2010; Breza, Kaur, and Shamdasani, 2018; Cohn, Fehr, Herrmann, and Schneider, 2014). Finally, Garicano, Palacios-Huerta, and Prendergast (2005) show that football referees systematically favour home teams. Through the social image treat-

ment, we contribute to empirical behavioural literature on the role of social pressure. We build on these results by showing that such social image concerns can bias individual decision making, and that this appears to improve the quality of such decisions. This result provides further empirical support for the finding in DellaVigna et al. (2012) and Dellavigna et al. (2016) that social image concerns matter even if a repeat interaction between individuals is unlikely.

Second, we contribute to the literature on organisational messaging. For example, Khan (2020) shows that promoting an organisation's mission can provide a valuable alternative to financial incentives to promote effort amongst public sector employees. Ashraf, Bandiera, and Jack (2014) similarly shows that non-financial incentives can be effective in motivating public employees in settings where financial awards are not feasible. There is also some related evidence in the management literature on corporate social responsibility being used as a tool to increase employee engagement (Flammer and Luo, 2017). We contribute to this literature by providing some of the first evidence on the role of communicating organisational values in promoting equal opportunity in business plan competitions.

Our findings offer new insights into the dynamics of institutional decision-making, particularly concerning the promotion of inclusivity through organisational messaging and social pressure. The remainder of this paper is structured as follows: Section 4.2 sets out a theoretical model to motivate the experiment, after which section 4.3 details the experimental design, and section 4.4 details the main results from the experiment. To better understand what drives our result, we implement a number of machine learning methods in section 4.5. We discuss our results in section 4.6.

4.2 Theoretical motivation

In this section, we present a stylised theoretical framework to capture the basic trade-offs that organisations face when they design assessment processes. On the one hand, the organisation wants decision-makers to use their discretion and good judgement in assessing candidates; on the other hand, the organisation cares about some candidate

characteristics in particular – so does not want decision-makers purely to use their own personal and idiosyncratic values.

To capture this tension – and to consider potential organisational design options – we imagine (as in our experiment) that each judge must decide between a female and a male candidate. Each judge (j), when assessing a pair of candidates (p), receives two signals. Each signal relates to the performance of the female candidate relative to the male candidate: (i) ε_{jp}, an idiosyncratic signal reflecting the judge's 'personal' assessment of the pair, and (ii) x_{jp}, a signal reflective of the values of the organisation. We assume $\varepsilon_{jp} \sim \mathcal{N}(\mu_\varepsilon, 1)$ and $x_{jp} \sim \mathcal{N}(\mu_x, 1)$ – so that μ_ε captures the judge's personal preference for voting in favour of women, and μ_x reflects the judge's perception of the organisational preference for favouring women. Judge j votes for the female candidate in pair p if the judge perceives a net positive signal in favour of the female. For tractability, we express the convex combination of signals using the unit circle:[40]

$$s_{jp}(\beta) \equiv \sin(\beta) \cdot x_{jp} + \cos(\beta) \cdot \varepsilon_{jp}, \tag{4.17}$$

where judge j votes for the female candidate in pair p if and only if $s_{jp} > 0$.

We assume that a judge who is invited simply to express an individual opinion will set $\beta = 0$; that is, the judge will vote based upon ε_{jp}. However, a judge in a competition must decide in what capacity they are making recommendations: as an individual or as part of a larger organisation. A judge who cares about representing the organisation – in effect, wishing to adopt an identity of representing the organisation and its goals (Akerlof and Kranton, 2000) – will choose to rotate the signal towards x_{jp}. In our setting, we see the organisational value treatment as potentially increasing β directly by affecting the choice of identity.

Many organisational decisions have a strong element of social accountability: decision-makers are often required to justify their decisions among their peers and may have a preference for agreement (Asch, 1951) or to be able to justify their decisions in case they

[40] In particular, this trigonometric formulation provides an elegant way of ensuring that the resulting variance does not depend upon β.

dissent (Bursztyn, Egorov, Haaland, Rao, and Roth, 2023). We assume that the idiosyncratic signal ε_{jp} is drawn independently across judges; in contrast, we assume that x_{jp} represents some 'true' signal x_p with perception error ω_{jp}. This assumption reflects the notion that x_{jp} relates in some sense to the 'fundamentals' of a candidate, observable (albeit imperfectly) by other judges. We assume $x_p \sim \mathcal{N}(0,1)$ and $\omega_{jp} \sim \mathcal{N}(0,1)$. Again, for tractability, we use a unit circle (here, weighting with parameter γ, and imposing an intercept μ_x):

$$x_{jp} = \mu_x + \sin(\gamma) \cdot x_p + \cos(\gamma) \cdot \omega_{jp}. \tag{4.18}$$

Suppose that – as in our experiment – a judge faces the prospect of needing to justify her decision to two colleagues (judges k and l) after voting. This implies a second reason that organisational design features might increase the weight on organisational values: a judge facing social pressure may choose $\beta^* > 0$ in order to increase the probability of agreeing with peers and to make it easier to justify the decision. In particular, if we denote by Φ_3 the CDF of the trivariate Normal, the probability of judge j dissenting from both of her peers k and l is:[41]

$$\begin{aligned}\Pr(\text{judge } i \text{ dissents}) = &\underbrace{\Phi_3\left(\mu(\beta^*)\cdot\begin{pmatrix}-1\\1\\1\end{pmatrix}, \begin{pmatrix}1 & -C(\beta^*) & -C(\beta^*)\\ -C(\beta^*) & 1 & C(\beta^*)\\ -C(\beta^*) & C(\beta^*) & 1\end{pmatrix}\right)}_{\text{dissent in favour of the male}}\\ &+\underbrace{\Phi_3\left(\mu(\beta^*)\begin{pmatrix}1\\-1\\-1\end{pmatrix}, \begin{pmatrix}1 & -C(\beta^*) & -C(\beta^*)\\ -C(\beta^*) & 1 & C(\beta^*)\\ -C(\beta^*) & C(\beta^*) & 1\end{pmatrix}\right)}_{\text{dissent in favour of the female}},\end{aligned} \tag{4.19}$$

where $\mu(\beta^*) = \sin(\beta^*) \cdot \mu_x + \cos(\beta^*) \cdot \mu_\varepsilon$ and $C = \sin^2(\beta^*) \cdot \sin^2(\gamma)$.

[41] This follows from the fact that, under this signal structure, the joint distribution of the three judges' signals (each judge having chosen a common β^*) is: $(s_{jp}, s_{kp}, s_{lp})' \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\mu}^* = [\sin(\beta^*) \cdot \mu_x + \cos(\beta^*) \cdot \mu_\varepsilon] \cdot \mathbf{1}_3$ and the covariance matrix $\boldsymbol{\Sigma}^*$ has diagonal $\mathbf{1}_3$ and off-diagonal elements $\sin^2(\beta^*) \cdot \sin^2(\gamma)$.

Depending upon the values of μ_x, μ_ε and γ, social pressure can tilt the judges' incentives in different directions. To illustrate, consider several key cases. First, suppose that judges have a strong private preference in favour of men ($\mu_\varepsilon << 0$), and that the organisational preference is in favour of equality ($\mu_x = 0$). Suppose that the organisation-relevant signal has no common component ($\gamma = 0$). Then the organisational values treatment increases the probability of a woman winning, but reduces the probability of agreement among judges. The social image treatment itself has no effect, as $\beta = 0$ for individual judges, but reduces the effect of the organisational values treatment in the combined treatment. This happens because increasing the weight on the organisation-relevant signal now comes at the cost of a higher cost of disagreement, attenuating the effect of the organisational values treatment.

A second possibility is that judges again have a strong private preference in favour of men ($\mu_\varepsilon << 0$), and that the organisational preference is in favour of equality ($\mu_x = 0$). Now, instead suppose that the organisation-relevant signal has a common component ($\gamma > 0$). The effect of the organisational values treatment on the probability of female candidates winning remains positive, but the effect on coordination now depends on the relative magnitude of μ_ε and γ – if γ is sufficiently large relative to μ_ε, coordination increases, otherwise it falls. Similarly, the effect of the social image treatment on both the female candidate winning and on agreement among judges is now ambiguous depending on the relative magnitude of μ_ε and γ in the same way. The combined treatment's effect is with the social treatment either reinforcing or attenuating the organisational values treatment.

A third possibility is that both the private signals and the organisation-relevant signals are balanced between male and female candidates ($\mu_\varepsilon = \mu_x = 0$) – but that the organisation nonetheless favours particular commonly-observed candidate features (that is, $\gamma > 0$). In this case, none of the treatments changes the average probability of women winning. However, the three treatments might change *which* candidates win – in particular, by encouraging judges to focus on 'objective' (organisation-relevant) features of candidates, rather than relying on judges' own idiosyncratic impressions. Under the

organisational values treatment, this occurs because judges are nudged to focus on more 'organisation-relevant' features. Under the social image treatment, this occurs because a focus on the organisation-relevant signal increases the probability of concurrence between judges. In the combined treatment, these two effects can reinforce each other (if the separate treatment effects generate $\beta^* < \pi/2$); alternatively, if $\beta^* = \pi/2$, the individual treatments already maximise judges' incentive to rely upon the organisation-relevant signal, and there is no additional interaction effect.

This model motivates our experimental design, which allows us to flexibly assess the model's predictions. The experiment does not focus on a specific mechanism, but rather on how communicating organisational values and social image concerns shape how decision makers align individual decision making with organisational objectives.

4.3 Experimental design

In this section, we describe the setting for our experiment – a business plan competition in Addis Ababa. We go on to explain the four treatment conditions, and then describe the sample. Appendix 4.7.E provides detail on the pre-specified experimental design and randomisation.

4.3.1 Setting: A business plan competition

To implement our experiment, we ran a business plan competition among young professionals in Addis Ababa.[42] Competitors were invited to a studio to record a pitch for their business proposal, and these videos form the input for our experiment.[43] The competition took place in a league-style format: each competitor was placed in a pool of 10 candidates, with the winner of each pool receiving a cash prize of 50,000 Birr (equivalent to about USD 800 at the time of the experiment). Each pool of 10 candidates comprised five

[42] Specifically, young professionals were eligible for this competition based on their assessment by established firm managers in a separate field experiment: see Abebe (2020).

[43] Specifically, competitors were asked to prepare a three-minute pitch briefly introducing themselves and covering the following five points: (i) their business idea, (ii) their intended target market, (iii) their potential competition, (iv) their operations and (v) their cost of business.

female candidates and five male candidates. We asked the business plan judges to cast a series of binary votes – where each vote compared one female and one male candidate.[44]

4.3.2 Assessing the candidates: Four treatment conditions

Each judge was randomly assigned to one of four different conditions under which to assess the videos. Our experiment took place over a period of three weeks in the summer of 2022 – during which we ran two sessions per day, randomly varying which treatment arm was implemented in each session. Randomisation was done at the judge level: after agreeing to participate, each judge was assigned to one treatment condition (and invited to a specific session accordingly).

We grouped the assessments so that each pair of candidates was viewed by a total of 12 judges: three judges in each of the four treatment conditions. We randomised the order in which assessments took place (so that, for example, what is the first pair for one judge might be the sixth pair for a different judge). To avoid any learning or spillovers, judges were asked not to discuss the assessment of the competitors with other judges until they had finished all assessments.

We now describe each treatment condition in turn.

Condition 1: Control. Under the 'control condition', each judge assessed the videos individually, and without any messaging as to organisational purpose. Specifically, judges in the control group were invited to a venue in central Addis Ababa in groups of 12 to 15. All judges entered a single room in a classroom-style arrangement. Once all judges had arrived, the judges were shown two videos. The first video explained the competition and the judges' task (including details on prizes, and on how the competition winners would be determined). In the second video, a prominent Ethiopian businessman and former Olympian, Haile Gebrselassie, spoke briefly about the importance of the business plan competition in providing capital for Ethiopian entrepreneurs. Specifically, he said

[44] The remaining pairs – that is, the male-male pairs and the female-female pairs – were assessed by two human resources experts. In the interests of fairness across competitors, we weighted the votes such that the votes from the human resources experts had the same total weight as those of the judges.

the following:

> As you know, access to capital is limited for entrepreneurs in Ethiopia. This competition will provide an opportunity for entrepreneurs to access capital to start or grow their business. Your vote is important in deciding which individual will win the 50,000 Ethiopian birr prize; please consider your choices carefully.

After watching these videos, the judges started their assessments. These were done on tablet computers, with the assistance of facilitators. Specifically, each judge was played a series of video pairs; for each pair, the judge casts a vote for the candidate he or she recommends for the grant. Judges under the control condition knew that their peers were also assessing candidates – but judges were not told anything about which candidates their peers were assessing, and there was no subsequent discussion about any vote cast. Judges were only asked to, privately, provide some feedback to two of the candidates they assessed at the end of their assessments.

Condition 2: Organisational values condition. The 'organisational values' condition was designed to emphasise to judges that the organisers of the business plan competition particularly value the inclusion of female candidates. This condition differed from the control condition in just one respect. Specifically, the second video included an additional prompt on the importance of equal opportunity for the organisation – noting, in particular, the specific constraints that female entrepreneurs face in accessing capital in Ethiopia. We deliberately designed this as a light-touch treatment, acknowledging that judges need to consider a range of factors in assessing business plans. Specifically, the text is as follows (with the text in italics distinguishing the organisational values condition from the control condition):[45]

> As you know, access to capital is limited for entrepreneurs in Ethiopia. This competition will provide an opportunity for entrepreneurs to access capital

[45] The 'recent World Bank report' referred to in this text is the Ethiopia Gender Diagnostic Report (World Bank, 2022)

> to start or grow their business. *Considering equal opportunity: I realise you need to take into account a large number of factors when making your decision but would like you to keep in mind that when starting a business, female entrepreneurs face additional constraints due to lenders' biases. A recent World Bank report finds that male entrepreneurs are more likely to take out loans than female entrepreneurs. In terms of loan sizes, male entrepreneurs borrow about 50 percent more than female entrepreneurs. In this competition, we are committed to gender equality and want to promote male and female entrepreneurs equally.* Your vote is important in deciding which individual will win the 50,000 Ethiopian birr prize; please consider your choices carefully.

Condition 3: Social image. The 'social image' condition was designed to allow the possibility of social image concerns. We did this in several complementary respects. First, before watching the same two explanatory videos as judges in the control condition (and before starting their assessments), judges were asked to introduce themselves to the other judges – providing their name, company and position. Second, prior to each of their binary votes, each judge was shown the names and photographs of two other judges who would also be voting on the same pair of candidates. Third, at the conclusion of all voting, we randomly chose one of the pairwise votes; each judge then sat together with the other judges who assessed that pair (that is, the other two judges whose names and photographs had been shown before that particular vote). In their groups of three, judges explained and justified their votes to their peers, and jointly agreed on feedback for the two candidates. Judges under the 'social image' condition were told in advance, and before casting each vote, that this meeting would take place.

Condition 4: Combined condition. The final condition combined the features of the 'organisational values' condition and the 'social image' condition: this was designed to test whether communication of organisational values has a different impact when assessors are exposed to social image concerns. Specifically, this condition was identical to the 'social image' condition, but judges first watched the same videos as those in

the 'organisational values' condition.

4.3.3 Experimental participants: Professional human resource managers

We invited senior human resource managers from established Ethiopian firms to serve as judges.[46][47] We selected this sample of judges for several reasons. First, we had an established relationship with their firms. Second, they possess a clear understanding of the quality of young professionals in Ethiopia. Finally, they operate in an environment where social image concerns and organisational values are likely to be relevant.

Table 4.1 reports these summary statistics of the judges. Most judges are either the most senior manager or owner of the firm (33%) or human resources manager (32%). They are on average 41 years old, and have on average 20 years of professional experience, on average six of which are in their current position. The managers are highly educated, with 78% having a bachelor's degree and 76% having formal management education. Three quarters of the judges are male, one quarter are female. In the final column of Table 4.1 we report the p-value for a Wald test that the judges' characteristics are balanced across treatments, and find no evidence for imbalance for any of the characteristics.

The managers work for large firms with a median of 54 employees (mean 200). 95% of these firms are for-profit, 62% of these firms are private limited companies, and 16% are public limited companies. Most firms are located in the capital Addis Ababa, with a few based in the nearby cities of Adama and Bishoftu.

[46] These were drawn from a sample of managers that participated in a separate field experiment: see Abebe (2020).

[47] Appendix 4.7.E.4 details the invitation for judges, which is translated into Amharic before being shared with enumerators. Enumerators are told to closely stick to these scripts and not to give additional information discussing the experiment or how we expect judges to make their decisions.

Table 4.1: Judge-level summary statistics

	Overall	Control	Social Image	Org. Values	Both Treatments	p-value
Gender (1=male)	.74	.70	.80	.72	.74	.679
Has a bachelor's degree	.78	.77	.8	.81	.72	.672
Formal management education	.76	.77	.78	.72	.79	.804
Judge age	41	39	43	41	43	.246
Experience in current job	6.1	6.5	6.8	5.9	5.3	.474
Total experience (years)	20	19	20	20	20	.981
Position of manager						
Most senior manager/owner	33%	30%	42%	28%	32%	.318
Finance and administration	15%	16%	15%	13%	16%	.941
HR Manager	32%	39%	25%	33%	32%	.505
Other	20%	16%	17%	26%	21%	.43
Department of manager						
Human resources	40%	49%	34%	40%	39%	.409
Administration	34%	28%	42%	31%	35%	.373
Finance	13%	14%	10%	13%	14%	.914
Other	13%	9%	14%	17%	12%	.621
Number of judges	245	57	59	72	57	

Notes This table displays the average characteristics of judges in each of the treatment arms. The department of the manager is the main department in which the manager operates in their firm, the other category contains all departments in which fewer than 5% of managers work. Columns two to six report these characteristics by treatment, and column seven reports the p-value for a Wald test for the hypothesis that these means are equal across the treatment arms.

4.3.4 Additional assessment: Human resources experts

In addition to the experiment's primary assessments, we had four individuals independently evaluate all competition submissions. This group included two highly experienced enumerators and two senior HR managers. The enumerators focused on transcribing more 'factual' dimensions of the vignettes, such as how the competitors were dressed, how they composed themselves, and whether they addressed each required topic in their pitch.

The two senior managers, henceforth referred to as our 'HR Experts', first completed a shortened version of the enumerators' questionnaire. They then assessed the overall quality of the pitch on a scale from one to twenty and evaluated specific dimensions of the pitch, such as whether the competitors had a clear business concept and a strategy for growth. These latter questions were drawn from Fafchamps and Quinn (2018). The experts, who include the owner of an Ethiopian HR firm and the HR head of a large for-profit enterprise, were selected by our local partners for their in-depth understanding of the labour market and their ability to assess the quality and viability of business plans.

4.4 Results

To analyse our experiment, we run a series of OLS regressions, of the following form:[48]

$$\begin{aligned} y_{jp} = \quad & \beta_1 \cdot \texttt{Organisational\_Values}_j + \beta_2 \cdot \texttt{Social\_Image}_j \\ & + \beta_3 \cdot \texttt{Combined\_Treatment}_j + \mu_p + \varepsilon_{jp}, \end{aligned} \tag{4.20}$$

where j indexes judges and p indexes pairs of competing candidates. $\texttt{Organisational\_Values}_j$ is a dummy for judge j being assigned to the Organisational Values treatment (but not to the 'combined treatment'), $\texttt{Social\_Image}_j$ is a dummy for judge j being assigned to the Social Image treatment (but not to the 'combined treatment'), and $\texttt{Combined\_Treatment}_j$

[48] Appendix 4.7.B presents our pre-analysis plan along with the exact results from the specified regressions. In the main text, we highlight any regressions that were not pre-specified. While we had pre-specified analyses on the effect of gender composition of committees, the results show a clear null effect; see Table 4.6 for the regression results.

is a dummy for judge j being assigned to the combined treatment. We cluster errors at the grouped triplet level within a treatment throughout.[49]

4.4.1 Result 1: No effect on the probability of female candidates winning

We first estimate equation 4.20 using as the dependent variable a dummy for whether the judge voted for the female candidate rather than the male. We report results in Figure 4.1. The bottom panel of Figure 4.1 shows that, in the control group, judges have, on average, a 48% probability of voting for the female candidate; this barely differs at all between male and female judges. The top panel of Figure 4.1 shows no significant effect of any treatment. Given the lack of an evident bias towards male candidates in the control group, this is perhaps not surprising.

4.4.2 Result 2: Treatments increase agreement with human resources experts; this is significant for the 'organisational values' condition

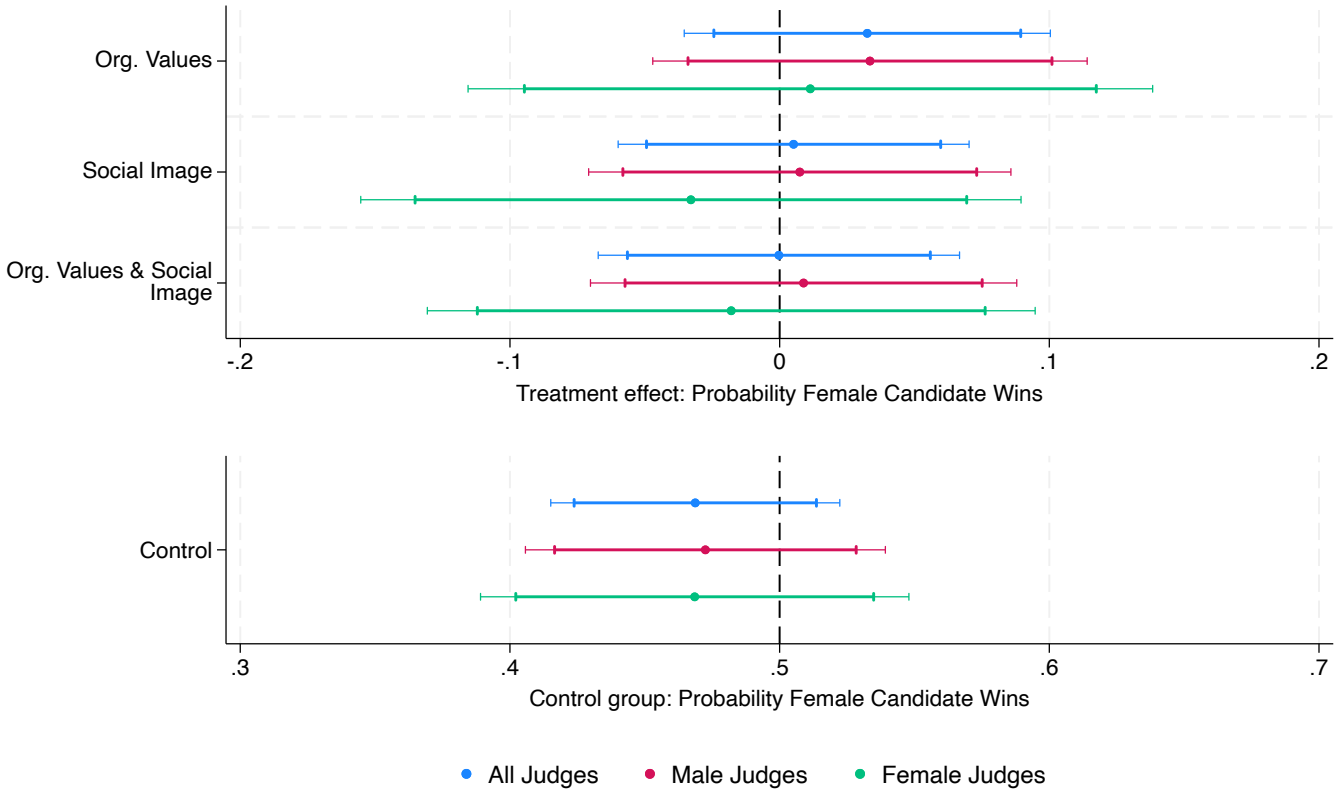
As noted, two human resources experts also assessed the videos. These scores were never shared with the judges; rather, the scores allow us to assess whether the treatments caused judges to vote for candidates whom experts consider to be better candidates. To do this, we estimate regression 4.20 using as the dependent variable a dummy for whether a judge voted for the candidate whom the human resources experts preferred.[50][51] The top panel in Figure 4.2 shows positive point estimates from all treatments on the probability of the expert's pick winning; this effect is significant only for the 'organisational values' condition.

[49] For one session, we had only male judges. We drop this session from the analysis as this violates the protocol for our assessments.

[50] Specifically, we define as the expert's pick the candidate with the higher average score across the two experts (with no expert pick in the case of average scores being equal).

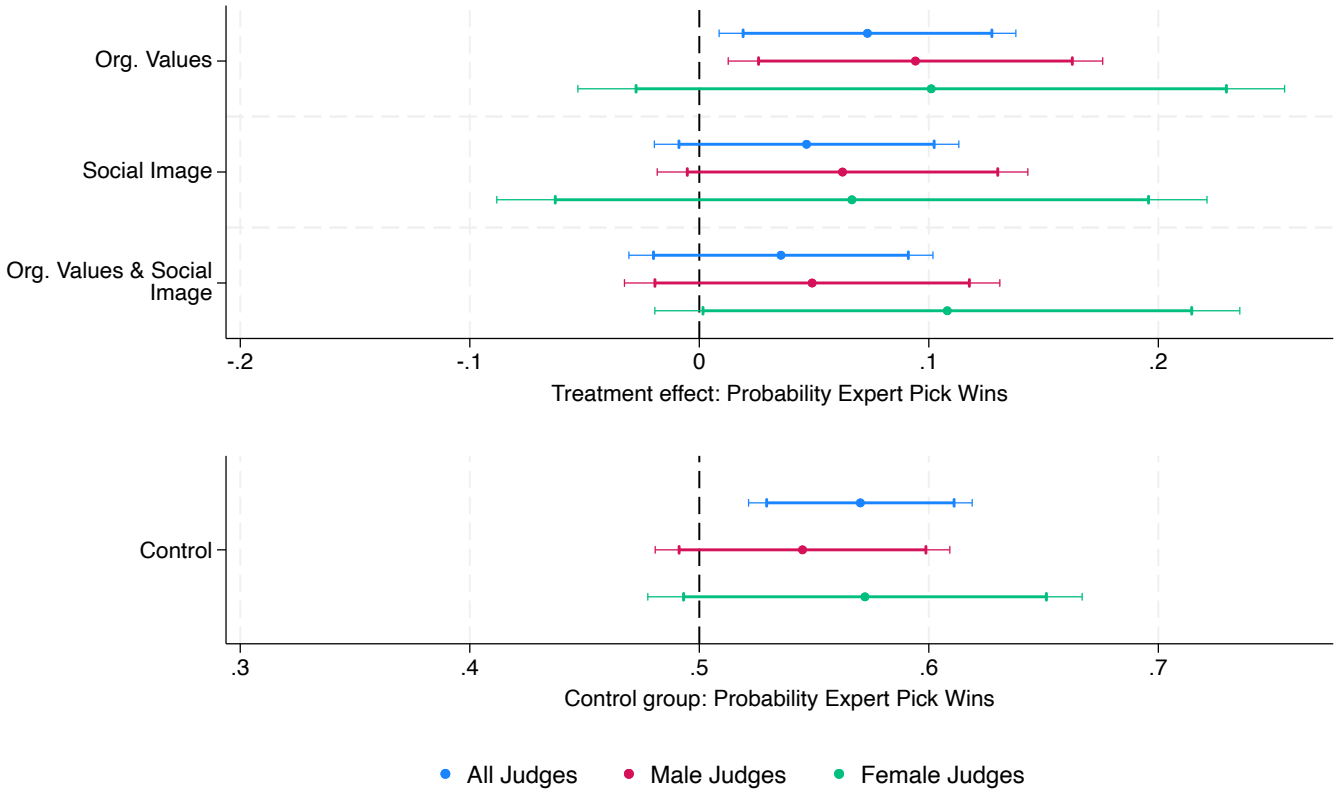
[51] This specification was not pre-specified

Figure 4.1: Effects on the probability of a female candidate winning



Notes This figure shows the effect of the different treatment conditions on the probability of voting for a female candidate. The bottom panel shows the probability of female candidates winning under the control condition; the top panel shows how this probability changes (that is, the treatment effects) for each of the three treatment conditions. We show effects for all judges, male judges and female judges. In each case, we show a point estimate and the 90% and 95% confidence intervals. Appendix Table 4.6 provides the corresponding numeric regression results.

Figure 4.2: Effects on agreement with expert assessments



Notes This figure shows the effect of the different treatment conditions on the probability of voting for a candidate who receives a higher average score from the expert assessors ('the expert's pick'). The bottom panel shows the probability of voting for the expert's pick; the top panel shows how this probability changes (that is, the treatment effects) for each of the three treatment conditions. We show effects for all judges, male judges and female judges. In each case, we show a point estimate and the 90% and 95% confidence intervals. Appendix Table 4.10 provides the corresponding numeric regression results.

4.4.3 Result 3: Treatments increase unanimity through coordination on candidates preferred by human resources experts

We now run a series of regressions to assess the effect of the treatments on the likelihood that the three judges assessing a given candidate pair reach the same decision (that is, unanimity). To do this, we define four separate dependent variables: (i) a dummy for unanimity, (ii) a dummy for unanimity in favour of the male candidate, (iii) a dummy for unanimity in favour of the female candidate, and (iv) a dummy for unanimity in favour of the candidate preferred by the human resources experts.[52]

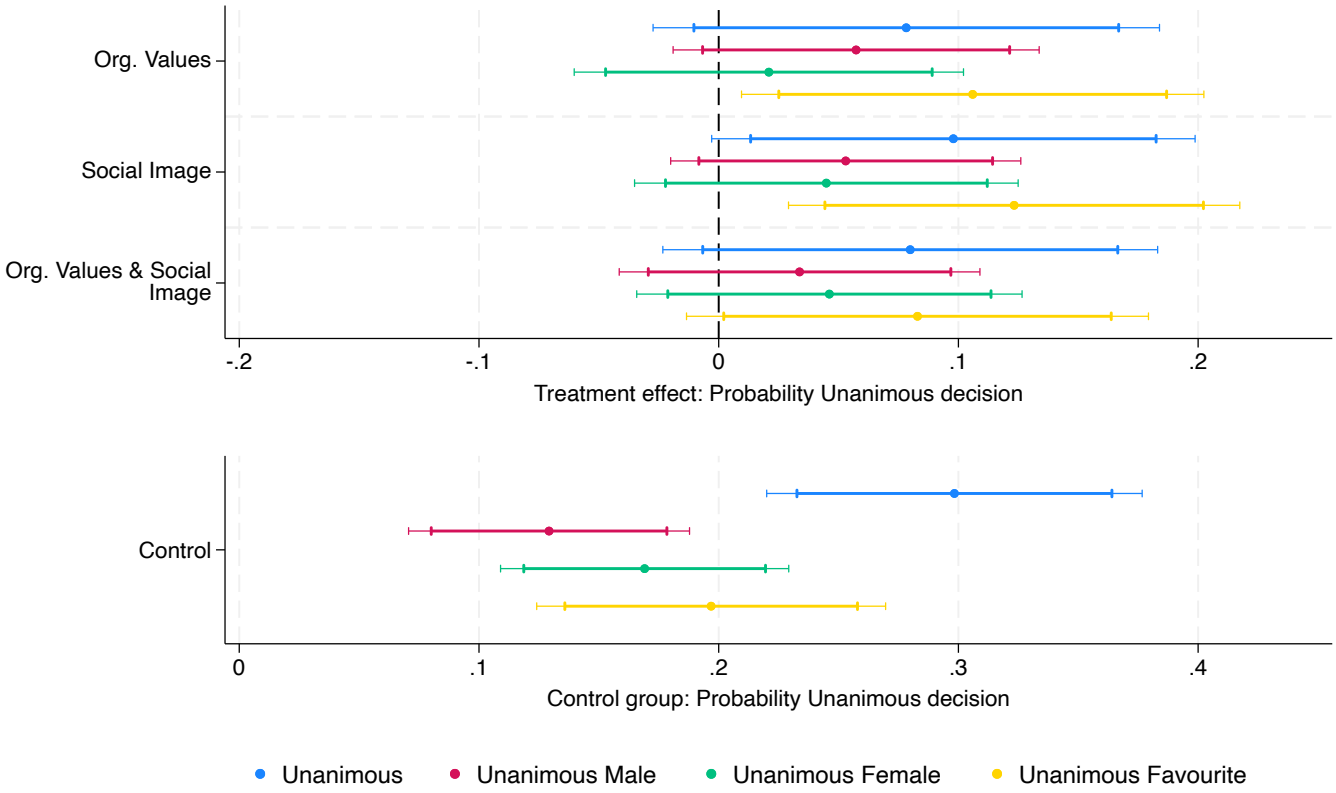
Figure 4.3 shows the results from these four regressions. We find positive point estimates for all treatments. This is particularly pronounced for the fourth outcome – unanimity in favour of the candidate preferred by the human resources experts. For this outcome, effects are significant across all treatments at the 90% confidence level. Moreover, the effect sizes are large. In the control group, the favorite wins unanimously 19% of the time (compared to the 12.5% that would be expected if each judge were simply to decide each decision by tossing a coin). This probability increases by about 10 percentage points in each of the three treatment conditions.

4.4.4 Result 4: Results 2 and 3 are driven by judges voting for the candidates whom experts strongly prefer

We now divide the sample into three groups, based on the terciles of the distribution of average expert scores within pairs. We then estimate equation 4.20 for these three subsamples, with the outcome being a dummy variable for whether the judge votes for the candidate preferred by the experts.[53] Figure 4.4 shows the results: for each treatment, the average treatment effect is over 10% for the top tercile in terms of score difference and close to zero for the other two terciles. In Figure 4.4, we repeat the exercise with

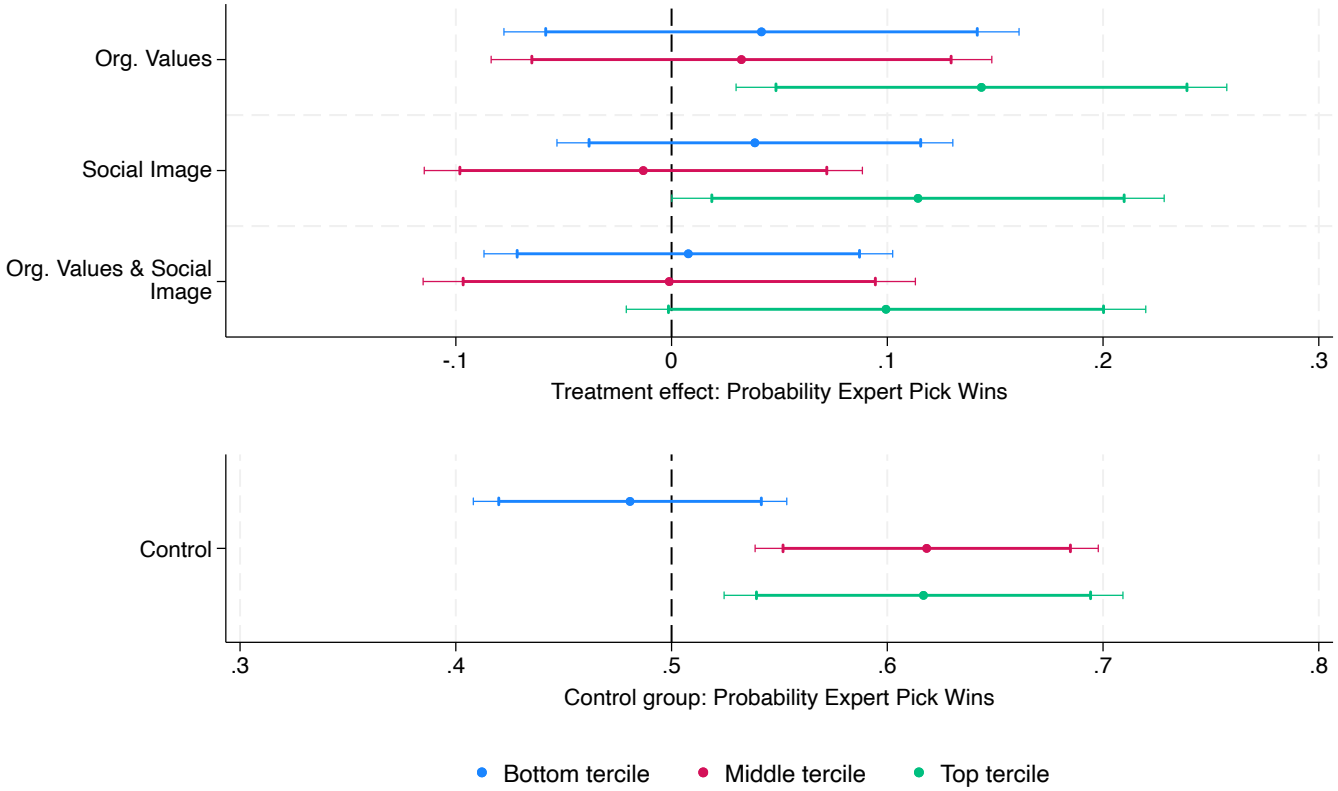
[52] Regression (iv) was not pre-specified
[53] This regression was not pre-specified

Figure 4.3: Effects on unanimity among grouped triplets of judges



Notes This figure shows the effect of the different treatment conditions on the probability of unanimity among the grouped triplets of judges who – within each treatment arm – assessed the same pair of candidates. We show effects (i) on unanimity, (ii) on unanimity in favour of the male candidate, (iii) on unanimity in favour of the female candidate, and (iv) on unanimity in favour of the candidate preferred by the human resources experts. In each case, we show a point estimate and the 90% and 95% confidence intervals. Appendix Table 4.8 provides the corresponding numeric regression results.

Figure 4.4: The effect of the treatments on the probability of the experts' pick winning by expert score difference.



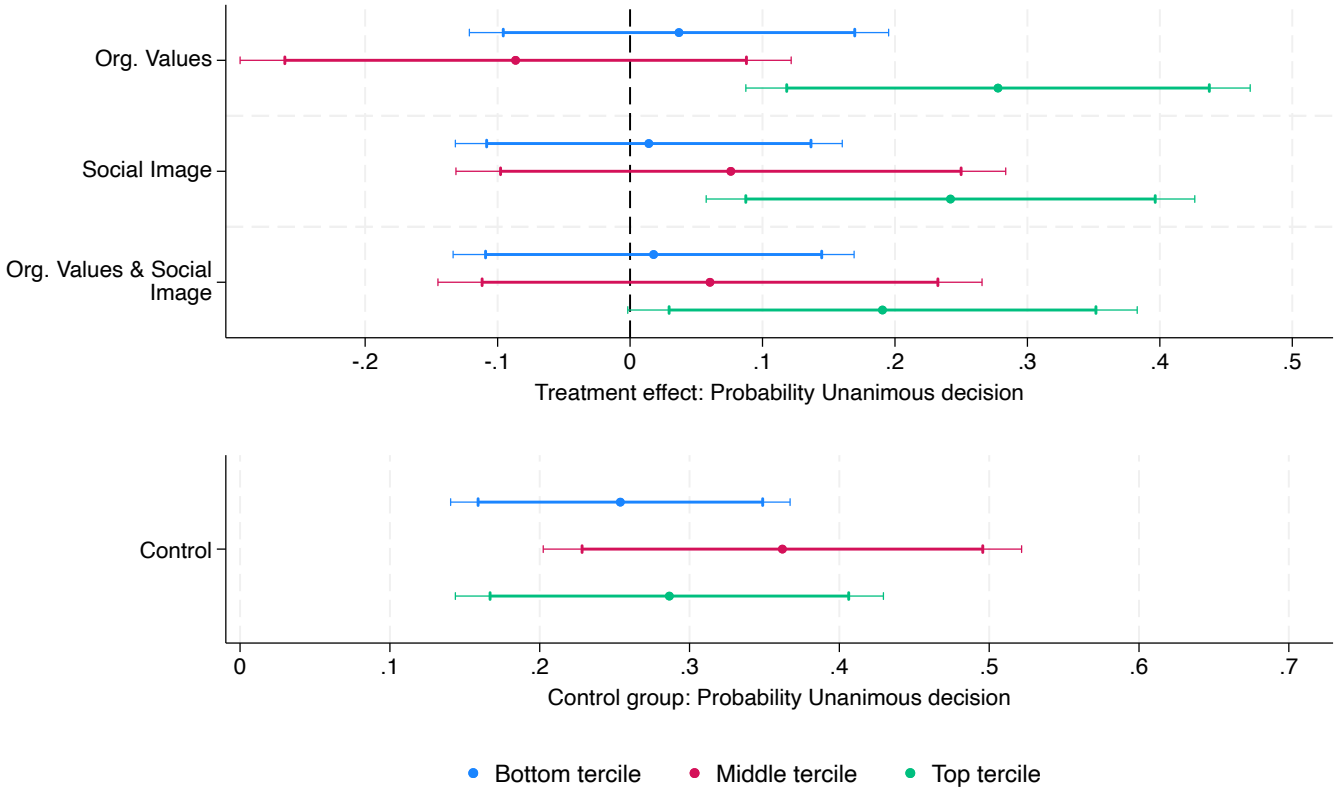
Notes This figure shows the effect of the different treatment conditions on the probability of the expert's pick winning. We show effects for three subsamples of the pairs of candidates in terms of the tercile of the absolute expert score difference for (i) the bottom tercile, (ii) the middle tercile, (iii) the top tercile. In each case, we show a point estimate and the 90% and 95% confidence intervals. Appendix Table 4.11 provides the corresponding numeric regression results.

the outcome being a dummy variable for unanimity among the grouped triplet. Here, the result is even more pronounced: the probability of unanimity increases by over 25 percentage points (on a control group mean of about 28%) for the top tercile in score difference.

4.5 Decision quality

The previous section showed that the treatments increase coordination among judges – and, in particular, that they do so by increasing the probability that judges vote for candidates who are preferred by human resources experts. This result is strongly suggestive that the treatments increase the quality of decisions – in the sense of preferring

Figure 4.5: The effect of the treatments on the probability of a unanimous decision by expert score difference.



Notes This figure shows the effect of the different treatment conditions on the probability of unanimity among the grouped triplets of judges who – within each treatment arm – assessed the same pair of candidates. We show effects for three subsamples of the pairs of candidates in terms of the tercile of the absolute expert score difference for (i) the bottom tercile, (ii) the middle tercile, (iii) the top tercile. In each case, we show a point estimate and the 90% and 95% confidence intervals. Appendix Table 4.12 provides the corresponding numeric regression results.

candidates on the basis of the quality of their business proposals.

We now take two complementary approaches to explore this hypothesis further. First, we implement two machine learning algorithms; these allow us to characterise heterogeneity in preferences among judges in the control group and heterogeneity in treatment effects. Second, we collect follow-up data on the competitors one year after the competition, in order to assess the predictive value of voting under the different treatment conditions.

4.5.1 Machine learning and heterogeneity

We use two machine learning approaches. First, we want to describe how predictable judges' choices are in each of the treatment conditions using observed characteristics. For this exercise, we use a random forest algorithm which we estimate separately for each of the treatment conditions. Second, we want to assess heterogeneous treatment effects, and specifically whether certain subsets of female candidates have particularly benefited from the treatments. To do this, we use the generic machine learning method of Chernozhukov et al. (2020).

For both algorithms, we use the same set of covariates. Specifically, we construct a rich set of measures of candidates' proposals and their presentations. We list these in detail in Appendix Table 4.4; in short, they comprise (i) detailed measures of the candidates' business proposals (including experts' assessments of those proposals), (ii) detailed measures of the justifications provided by the candidates in the presentations, (iii) detailed measures of the candidates personal appearance and (iv) based on a separate proprietary algorithm, scores reflecting the emotiveness of candidates' pitches based on their facial expressions.[54] In each case, we encode these characteristics at the pair level as the difference for each pair between the female and the male candidate; in each case, the outcome variable is the probability of the female candidate winning.[55]

[54] These include measures of how positive (or negative) the emotions displayed by a competitor are, and how engaged the competitor is during the pitch.

[55] All differences are normalised to a mean zero, variance one distribution. This means that a difference in the independent variables of 1 can be interpreted as a one standard deviation difference.

Random Forest. We run our random forest algorithm separately for each treatment arm; we then measure the 'out-of-bag' predictive accuracy. This measures how well the model predicts on data not used during its training (using only the samples left out in each bootstrap iteration). As the random forest algorithm imposes the same function $\hat{y} = f(x)$ for all judges in a treatment arm, a higher predictive accuracy implies a more consistent mapping from observable characteristics to votes across judges.

For control group judges, result 4 shows the dimension of quality captured by the experts' preferences is not strongly predictive of their decisions. This random forest algorithm allows us to assess whether these judges are coordinating on some other set of observable characteristics. If control judges were, for example, coordinating primarily on respondents' confidence and presentation skills, this would result in a high out-of-bag predictive accuracy for the control group; in contrast, if those judges were instead making decisions in a highly idiosyncratic way (for example, each judge fixating on a different candidate characteristic, or each judge acting almost as if flipping a coin), this would generate a low out-of-bag accuracy.

Running this algorithm, we find a far lower predictive accuracy for the control group relative to each of the treatment groups. The algorithm returns a 55% out-of-bag predictive accuracy for the control group, compared to 61-65% for each of the treatment groups.[56] This shows that the behaviour of the control group is highly idiosyncratic (we would expect a 50% out-of-bag accuracy for judges who make fully random decisions). Further – as our earlier Results 2, 3 and 4 indicated – it illustrates that each of the three treatments increases quite substantially the coordination on observable characteristics.[57]

Heterogeneous Treatment Effects. We implement Chernozhukov et al. (2020)'s generic machine learning algorithm to assess heterogeneity in treatment effects.[58] This approach first allows us to demonstrate that there is heterogeneity in the treatment effects, which can be captured using our chosen predictor when pooling all treatments. We

[56] This is respectively 61% for the organisational values treatment, 64% for the social image treatment and 65% for the combined treatment

[57] Note that the overall expert score is not included here in this predictive model, but that the individual scores for different dimensions of the quality of the pitch are quite predictive of the overall score.

[58] We use Welz, Alfons, Demirer, and Chernozhukov (2022)' R package to implement this analysis.

then show that the most positively affected group of female candidates receives almost half an additional vote per committee compared to the most negatively affected group ($p = 0.130$). The treatments appear to particularly benefit younger female candidates who are more engaged and give more specific answers, and to disadvantage more arrogant female candidates with a better appearance.

Throughout this analysis, we pool all treatment arms as "treated" individuals and compare these to the control group.[59] First, we estimate the Conditional Average Treatment Effects (CATEs) for all pairwise assessments. To do so, we use two machine learning models: A penalized regression using elastic net regularization, and a random forest algorithm.[60] We use the same set of variables described in Table 4.4.

Using these CATEs, we find that, as expected, there is no average treatment effect but significant heterogeneity in treatment effects. To do so, we run the following regression, where D is treatment status, $p(Z)$ is 75% as each individual has a 75% chance of being assigned to a treatment, and $S(Z)$ is the conditional average treatment effect.

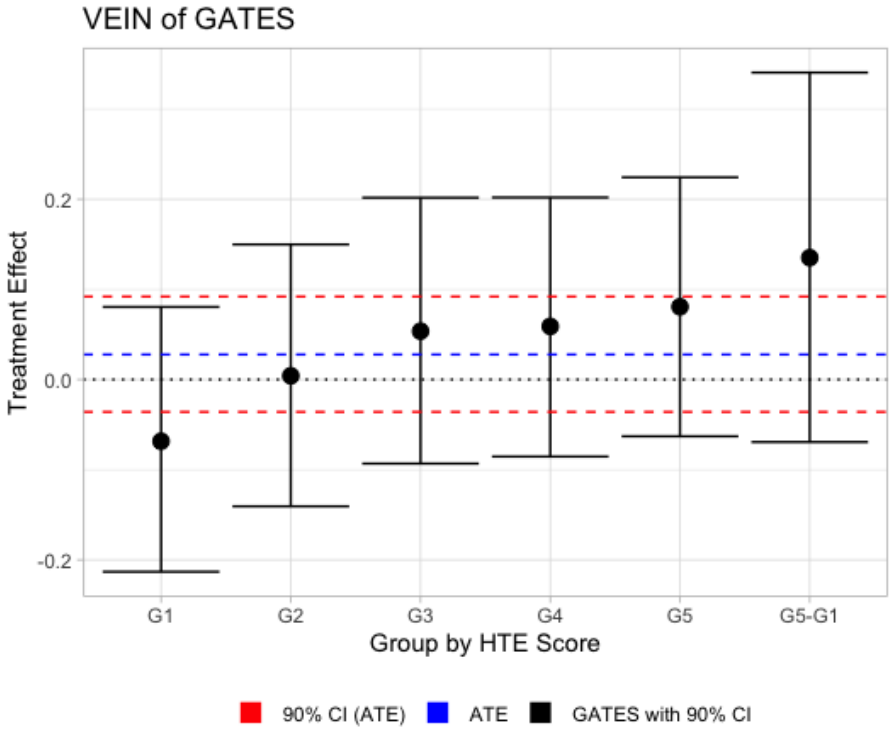
$$Y = \beta_1 (D - p(Z)) + \beta_2 ((D - p(Z)) (S(Z) - \mathbb{E}[S])) + \epsilon$$

The null hypothesis that $\beta_2 = 0$ tests whether there is heterogeneity in treatment effects and that S(Z) is a relevant predictor. We estimate $\hat{\beta}_2 = 0.15$ with a p-value of 0.13. We take this as weak evidence for the existence of heterogeneity in treatment effects.

Next, we look at the Grouped Average Treatment Effects (GATES). We divide the estimated CATEs into groups based on quintile cutoffs. We then assess what the difference in treatment effect is between the most (`G5`) and least (`G1`) affected group of candidates. Figure 4.6 shows the results of this analysis, which shows that the most affected group gets around 0.14 additional votes per judge. Note that throughout, group one is the group of pairs with the most *negatively* affected female candidates, and group five is the group of pairs with the most *positively* affected female candidates.

[59] Appendix B details the results of individual treatment arms compared to the control group.
[60] Using respectively the `cv_glmnet` learner and the `ranger` learner with 500 trees from the `mlr3` package

Figure 4.6: Group average treatment effects



Notes This figure shows the group average treatment effects of the five groups of pairwise assessments generated based on the conditional average treatment effects. It demonstrates that the most positively affected group of female candidates (`G5`) has a 13 percentage point (`G5-G1`) greater increase in the probability of being recommended than the least positively affected group (`G1`)

Finally, we implement classification analysis to assess the average characteristics of the most versus least positively affected group. Table 4.2 shows the results of this analysis. Female competitors in the most affected group are more likely to mention their family, teamwork and to give specific examples. They were less likely to be perceived arrogant, were perceived as less appropriately dressed and were on average younger.

Table 4.2: Classification Analysis: What do the groups look like

Variable	Mean group 1	Mean Group 5	Difference mean	p-value
Mentions family	-0.162	0.166	0.327	0.000
Average engagement	-0.154	0.148	0.299	0.000
Mentions teamwork	-0.149	0.106	0.230	0.003
Gives specific examples	-0.101	0.124	0.214	0.007
Arrogance	0.125	-0.066	-0.203	0.010
Mentions market	0.057	-0.145	-0.209	0.007
Appropriate appearance	0.074	-0.159	-0.262	0.001
Age	0.152	-0.127	-0.265	0.000
Care of outfit	0.148	-0.137	-0.276	0.001
General appearance	0.149	-0.147	-0.276	0.000

Notes This table depicts the classification analysis. The mean group 1 is the average difference between the female and male candidates' score in the most negatively affected group, mean group five is this average for the most affected group. The difference mean is the difference in these two values, and the p-value is the p-value for the hypothesis that this mean is zero. These are the variables returned for the CLAN at the 1% level.

4.5.2 Long-term labour market outcomes

Around 18 months following the competition, we collect follow-up data to assess what happened to the candidates after the competition. We collect data from 89 of the 100 candidates on their employment status, labour market outcomes, and some questions relating to their current business (plans). To understand whether judges recommended

different types of candidates across treatments, we run the following regression pooling the treatments, where the observation is the labour market outcomes of individual i:

$$\begin{aligned} Y_i = &\alpha + \beta_1 \texttt{Score\_Control}_i + \beta_2 \texttt{Score\_Organisational\_Values}_i \cdot \\ &\beta_3 \texttt{Score\_Social\_Pressure}_i \cdot \beta_4 \texttt{Score\_Both\_treatments}_i \cdot \\ &+ \zeta \texttt{Winner}_i + \varepsilon_i \end{aligned}$$

Where `Score_X` is the number of pairwise comparisons a candidate won in treatment arm X and `Winner` is an indicator for having won the prize. Table 4.3 describes the results from this regression,[61] including in Y an indicator for transitioning out of self-employment, self-employment, wage-employment, total income, total investment in their business since the competition, the number of steps they have taken to establish a business and whether or not they have taken out a business loan since starting the business. These measures aim to broadly capture the labour-market outcomes of the young professionals following the competition.

[61] We exclude the transition into self-employment as a dependent variable due to the small number of respondents that do so.

Table 4.3: Long-term outcomes and competition performance

	Transition into self-employment	Self-employed	Wage-employed	Total income (ETB1000)	Total investment	Total steps taken	Has loan
	b/se	b/se	b/se	b/se	b/se	b/se	b/se
Score control	0.026*	-0.005	0.004	0.079	0.149*	0.057	-0.023
	(0.01)	(0.02)	(0.02)	(0.06)	(0.08)	(0.22)	(0.02)
Score org.	0.006	0.018	0.007	0.048	0.148	0.123	0.030
values	(0.02)	(0.02)	(0.02)	(0.10)	(0.16)	(0.27)	(0.02)
Score social	-0.039**	-0.045*	-0.002	0.024	0.005	-0.170	-0.014
image	(0.02)	(0.02)	(0.02)	(0.08)	(0.12)	(0.31)	(0.02)
Score both	0.007	0.018	-0.005	0.052	-0.003	0.009	-0.025
treatments	(0.01)	(0.02)	(0.02)	(0.09)	(0.12)	(0.30)	(0.02)
Winner	0.048	0.206	0.125	0.224	0.202	0.766	0.407**
	(0.13)	(0.20)	(0.16)	(0.76)	(0.99)	(2.23)	(0.20)
Control mean	0.134	0.457	0.706	2.057	10.172	6.041	0.549
N	89	89	89	89	89	89	89

Notes This table describes the relationship between labor market outcomes and candidate performance in each of the treatment arms, i.e. control, organisational values, the social image treatment, and both treatments which is the group receiving both the organisational values and social image treatment. The transition into self-employment is a dummy equal to one if respondents was not in self-employment before but is after the competition. Self-employed and wage-employed are dummies. Total income is monthly income out of any source. Total investment is the amount the respondent has invested in their business since the competition, this is implemented as a Poisson regression to reduce the influence of outliers. Total steps taken is the number of steps (from a fixed list) a respondent has taken starting a business. Has loan is an indicator for whether the respondent has a business loan.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The results, presented in Table 4.3, show that overall, performance in the competition has limited predictive power for long-term labour market outcomes. This finding aligns with previous studies, such as McKenzie and Sansone (2019), which suggest that accurately predicting the future success of aspiring entrepreneurs based on competition performance is inherently difficult.

Despite the limited predictive power across most outcomes, there are a few notable patterns. Candidates who performed well in the control treatment (i.e. received high scores in the absence of any organizational or social pressure) are significantly more likely to transition into self-employment. Interestingly, those who scored well amongst control group judges also tended to invest more in their businesses. This suggests that the control group judges were slightly better at identifying candidates likely to grow their businesses. In contrast, candidates who performed well amongst judges in the social-image treatment are less likely to transition into self-employment and to be in self-employment.

Overall, the results indicate that while competition performance may not be a strong predictor of long-term success in various labour market outcomes, certain treatment-specific effects do emerge. High performers in the control group appear to have moved away from self-employment while those remaining invest relatively more in business activities. Conversely, those who do well in the social-image treatment are more likely to remain self-employed, though their ventures are less financially successful. Given the large number of regressions done here, these results are unlikely to survive rigorous multiple-hypothesis testing. In Appendix Table 4.13 we run this same specification with a dummy for above-average performance within the competition as dependent variable with comparable results.

4.6 Discussion and Conclusion

This study provides new insights into the interaction between organisational messaging and social image concerns within the context of institutional decision-making. Our results suggest that, while neither the organisational values nor the social image treatment significantly affected the likelihood of female candidates winning a business plan compe-

tition, these institutional interventions did enhance the overall quality of decisions. This is evidenced by the increased alignment of judges' votes with expert assessments, particularly in the treatment groups exposed to organisational values and social image concerns. These findings underscore the potential for light-touch institutional interventions to promote decision-making that is more in line with organisational goals, particularly when these goals pertain to inclusivity and equal opportunity.

Interestingly, our results indicate that, while the treatments improved decision quality as reflected by agreement with experts, they did not directly increase the probability of female candidates winning. This outcome suggests that, at least in the context of this competition, gender bias was not a major factor in the control group, where male and female candidates were equally likely to win. Thus, interventions aimed at promoting gender equity may instead play a role in improving decision quality rather than directly addressing gender disparities in outcomes.

However, the absence of a significant effect on the probability of female candidates winning also raises important questions about the limitations of such interventions. While organisational messaging and social image concerns can nudge individuals toward more consistent decisions and greater agreement among decision-makers, deeper structural changes may be necessary to address systemic biases and promote greater inclusivity. Future research could explore the interaction between institutional interventions and more robust mechanisms, such as financial incentives or longer-term organisational commitments to diversity, to examine how these factors might work together to drive more inclusive outcomes.

4.7 Appendices

4.7.A Variables

Table 4.4: Observable characteristics of submissions

Pair Level	
ASPIRE	**Expert**
Score for business concept	Likert 1-5
Score for understanding market	Likert 1-5
Score for growth strategy	Likert 1-5
Score for presentation	Likert 1-5
Score for business plan	Likert 1-5
Score for business sense	Likert 1-5
Overall score	Discrete 1-20***
Required Content Included	**Expert & Enumerator**
Introduces themselves	Binary* ***
Mentions business idea	Binary**
Mentions target market	Binary
Mentions competition	Binary
Mentions operations	Binary
Mentions business costs	Binary
Other Expert Questions	**Expert**
Is appearance appropriate	Likert 1-4
Is the competitor confident	Likert 1-4
Does the competitor seem arrogant	Likert 1-4
How certain or convincing?	Likert 1-4
Planned location of firm	Indicator***
Start or expand firm	Binary***
Other Enumerator Questions	**Enumerator**
Ethnicity	Indicator
Gives specific examples	Binary
Is the competitor dressed formally	Likert 1-4
Did the competitor take care of their outfit	Likert 1-4
The competitor's general appearance is appropriate	Likert 1-4
Is the competitor confident	Likert 1-4
Does the competitor seem arrogant	Likert 1-4
Does the competitor give examples of teamwork	Indicator
Does the competitor mention their family	Indicator
Does the competitor thank others	Indicator
How old do you think the competitor is?	Discrete
Emotions	**Proprietary Algorithm**
Sum of positive and negative emotions	0-100
Sum of absolute positive and negative emotions	0-100
Duration of the submission	Discrete

* No variation, ** Almost no variation, *** Not included for ML algorithms.

Table 4.5: Judge characteristics

Judge Level	
Judge characteristics	**Scale/Type**
Hostile sexism (HS)	Continuous
Benevolent sexism (BS)	Continuous
Judge age	Discrete
Judge gender	Binary
Years of experience	Discrete
Industry of firm	Indicator
Position at firm	Indicator
Questions about the experiment	**Scale/Type**
Extent to which judge cares about feedback committee (careFeedback)	Likert
Extent to which judge considers votes of others (considerOthers)	Likert
Select up to 5 judges important for perception (importanceOthers)	Indicator
Importance of characteristics for proposal quality	
Good concept (Concept)	Likert
Understanding market (Market)	Likert
Strategy for growth (Growth)	Likert
Financial plan (Finance)	Likert
Business plan (Plan)	Likert
Presentation skill (Present)	Likert
Business sense (Sense)	Likert

Ninety-seven judges listed all these characteristics as 'Strongly agree' when asked about their importance.

4.7.B Results pre-specified regressions

4.7.B.1 The effect of the treatments on female candidates performance

We examine the effect of the three treatment arms using the following regression, i.e. regression 4.20 from the main paper:

$$\text{WomanWins}_{jcp} = \alpha + \beta_1 \text{Info}_j + \beta_2 \text{Comm}_j + \beta_3 \text{Info}_j \cdot \text{Comm}_j + \mu_p + \varepsilon_{jcp} \tag{4.21}$$

Stata: `reghdfe womanWins info comm infoComm, absorb(pair) vce(cluster judge)`

The variables are defined as follows:

- WomanWins_{jcp}: Judge j on committee c votes for the woman in pair p.

- Info_j: Judge j has the information treatment (set to 1 for treatments 1 and 3).
- Comm_j: Judge j is on a committee (set to 1 for treatments 2 and 3).
- μ_p: Pair fixed effects.
- ε_{jp}: Standard errors are clustered at the judge level.

For our secondary hypothesis we do subgroup analysis by gender. We look at the treatment effect for male and female judges separately using the specification in equation 4.20. We run regression 4.20 using OLS on the sub-sample of male and female judges and then compare the coefficients testing the same set of hypotheses for the comparison of the two sets of parameters.

$$\text{WomanWins}_{jcp} = \alpha + \beta_1 \text{Info}_j + \beta_2 \text{Comm}_j + \beta_3 \text{Info}_j \cdot \text{Comm}_j + \mu_p + \varepsilon_{jcp} \tag{4.22}$$

Stata: `reghdfe womanWins info comm infoComm if gender == 1, absorb(pair) vce(cluster judge)`

$$\text{WomanWins}_{jcp} = \alpha + \beta_1 \text{Info}_j + \beta_2 \text{Comm}_j + \beta_3 \text{Info}_j \cdot \text{Comm}_j + \mu_p + \varepsilon_{jcp} \tag{4.23}$$

Stata: `reghdfe womanWins info comm infoComm if gender == 2, absorb(pair) vce(cluster judge)`

Next we run the following regression to study the effect of committee composition, following regression 4.28 from the main paper.

$$\text{WomanWins}_{jcp} = \alpha + \beta_1 \text{Info}_j + \beta_2 \text{Woman}_c + \beta_3 \text{Info}_j \cdot \text{Woman}_c + \mu_p + \varepsilon_{jp}, \tag{4.24}$$

where $\text{Woman}_c = 1$ if there is a woman on the committee.

Stata: `reghdfe womanWins info woman infoWoman if comm==1 & gender ==1, absorb(pair) vce(cluster judge)`[62]

The results from these regressions are presented in Table 4.6.

[62] The condition & gender == 1 was not pre-specified erroneously. This means we focus on the effect of having a female committee member on male committee members.

Table 4.6: The effects of treatments and committee composition on the probability of voting for a female candidate

	Dependent variable: Female candidate wins			
Sample	All judges	Male judges	Female judges	Male judges
	b/se	b/se	b/se	b/se
Org. Values	0.033	0.033	0.011	0.108
	(0.03)	(0.04)	(0.06)	(0.07)
Social Image	0.005	0.007	-0.033	0.052
	(0.03)	(0.04)	(0.06)	(0.07)
Org. Values	-0.000	0.009	-0.018	0.040
& Social Image	(0.03)	(0.04)	(0.06)	(0.07)
female member				0.055
				(0.07)
Org. Values × female member				-0.142
				(0.09)
Social Image × female member				-0.064
				(0.08)
Org. Values & Social Image				-0.037
× female member				(0.09)
Control mean	0.469	0.472	0.469	0.432
N	2514	1844	628	1844

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.7.B.2 The effect of the treatments on the probability of dissent

Finally, we examine the outcomes at a committee level, examining the probability of a unanimous decision for each of the treatment arms. We first run the same regression as in equation 4.20, but now at a committee level with as dependent variable whether or not the committee's decisions was unanimous.

We then run the same regressions with as outcome variables whether or not the committee's decision was unanimously for a male candidate, and whether or not the committee's decision was unanimously for a female candidate. These regressions help us examine the channels through which the treatments affect judges' decisions. (For example, a high value for the committee treatment in the regression with as dependent variable UnanimousWomanWins would indicate judges in particular do not want to be found to be the only one dissenting by voting for the male candidate when the other judges' are expected to vote for the female candidate.)

We run the following three regressions:

$$\text{Unanimous}_{cp} = \alpha + \beta_1 \text{Info}_c + \beta_2 \text{Comm}_c + \beta_3 \text{Info}_c \cdot \text{Comm}_c + \mu_p + \varepsilon_{cp} \tag{4.25}$$

Stata: `reghdfe unanimous info comm infoComm, absorb(pair) vce(cluster committee)`

$$\text{UnanimousWomanWins}_{cp} = \alpha + \beta_1 \text{Info}_c + \beta_2 \text{Comm}_c + \beta_3 \text{Info}_c \cdot \text{Comm}_c + \mu_p + \varepsilon_{cp} \tag{4.26}$$

Stata: `reghdfe unanimousWomanWins info comm infoComm,`

```
absorb(pair) vce(cluster committee)
```

$$\text{UnanimousManWins}_{cp} = \alpha + \beta_1 \text{Info}_c + \beta_2 \text{Comm}_c + \beta_3 \text{Info}_c \cdot \text{Comm}_c + \mu_p + \varepsilon_{cp} \quad (4.27)$$

Stata: `reghdfe unanimousManWins info comm infoComm, absorb(pair) vce(cluster committee)`

Table 4.7: The effects of treatments on making a unanimous decision

	Unanimous b/se	Unanimous woman wins b/se	Unanimous man wins b/se	Unanimous expert pick wins b/se
Org. Values	0.078	0.057	0.021	0.106**
	(0.05)	(0.04)	(0.04)	(0.05)
Social Image	0.098*	0.053	0.045	0.123**
	(0.05)	(0.04)	(0.04)	(0.05)
Org. Values &	0.080	0.034	0.046	0.083*
Social Image	(0.05)	(0.04)	(0.04)	(0.05)
Control mean	0.298	0.129	0.169	0.197
N	823	823	823	762

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.7.B.3 The effect of committee composition on the probability of dissent

Finally, for the subsample of judges on committees, we test whether committee composition affects the probability of dissent at a committee level. To do this, we use the same dependent variables as in regression 4.28, but now use as a dependent variable the probability of a unanimous decision.

$$\text{Unanimous}_{cp} = \alpha + \beta_1 \text{Info}_c + \beta_2 \text{Woman}_c + \beta_3 \text{Info}_c \cdot \text{Woman}_c + \mu_p + \varepsilon_{cp} \quad (4.28)$$

Stata: `reghdfe unanimous info woman infoWoman if comm==1, absorb(pair) vce(cluster committee)`

$$\text{UnanimousWomanWins}_{cp} = \alpha + \beta_1 \text{Info}_c + \beta_2 \text{Comm}_c + \beta_3 \text{Info}_c \cdot \text{Comm}_c + \mu_p + \varepsilon_{cp} \quad (4.29)$$

Stata: `reghdfe unanimousWomanWins info woman infoWoman if comm==1, absorb(pair) vce(cluster committee)`

$$\text{UnanimousManWins}_{cp} = \alpha + \beta_1 \text{Info}_c + \beta_2 \text{Comm}_c + \beta_3 \text{Info}_c \cdot \text{Comm}_c + \mu_p + \varepsilon_{cp} \quad (4.30)$$

Stata: `reghdfe unanimousManWins info woman infoWoman if comm==1, absorb(pair) vce(cluster committee)`

Table 4.8: The effects of the information treatment and committee composition on making a unanimous decision

	Unanimous b/se	Unanimous woman wins b/se	Unanimous man wins b/se
Org. Values	-0.009	-0.020	0.011
	(0.09)	(0.06)	(0.08)
female member	0.137*	0.073	0.064
	(0.08)	(0.06)	(0.06)
Org. Values × female member1	-0.028	0.019	-0.046
	(0.13)	(0.09)	(0.10)
Control mean	0.000	0.000	0.000
N	444	444	444

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.7.C Numerical representation main results

The tables in this section present the regression results also included in the main empirical analysis. All regressions are run both without controls, and controlling for the share of female judges attending a session. The latter control is included as the share of female judges if more variable than expected, but this does not meaningfully affect regression results.

Table 4.9: The effects of treatments on the probability of voting for a female candidate

	All judges		Male judges		Female judges	
	b/se	b/se	b/se	b/se	b/se	b/se
Org. Values	0.033	0.028	0.033	0.036	0.011	-0.012
	(0.03)	(0.03)	(0.04)	(0.04)	(0.06)	(0.05)
Social Image	0.005	-0.008	0.007	-0.002	-0.033	-0.030
	(0.03)	(0.03)	(0.04)	(0.04)	(0.06)	(0.08)
Org. Values &	-0.000	-0.007	0.009	0.015	-0.018	-0.029
Social Image	(0.03)	(0.03)	(0.04)	(0.04)	(0.06)	(0.06)
Controls	No	Yes	No	Yes	No	Yes
Control mean	0.469	0.476	0.472	0.472	0.469	0.472
N	2514	2514	1844	1844	628	628

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.10: The effects of treatments on the probability of voting for an expert-favoured candidate

	All judges		Male judges		Female judges	
	b/se	b/se	b/se	b/se	b/se	b/se
Org. Values	0.073**	0.063*	0.094**	0.088**	0.101	0.039
	(0.03)	(0.03)	(0.04)	(0.04)	(0.08)	(0.08)
Social Image	0.047	0.037	0.062	0.061	0.066	0.050
	(0.03)	(0.03)	(0.04)	(0.04)	(0.08)	(0.07)
Org. Values &	0.036	0.021	0.049	0.046	0.108*	0.054
Social Image	(0.03)	(0.03)	(0.04)	(0.04)	(0.06)	(0.06)
Controls	No	Yes	No	Yes	No	Yes
Control mean	0.570	0.580	0.545	0.549	0.572	0.617
N	2331	2331	1708	1708	583	583

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.11: The effects on probability of voting for the experts' favourite by score difference

	Dependent variable: Favourite Wins					
	b/se	b/se	b/se	b/se	b/se	b/se
Org. Values	0.042	0.041	0.032	0.010	0.144**	0.140**
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Social Image	0.039	0.048	-0.013	-0.040	0.114**	0.102*
	(0.05)	(0.04)	(0.05)	(0.05)	(0.06)	(0.06)
Org. Values &	0.008	-0.005	-0.001	-0.029	0.099	0.090
Social Image	(0.05)	(0.05)	(0.06)	(0.06)	(0.06)	(0.06)
Controls	No	Yes	No	Yes	No	Yes
Control mean	0.481	0.480	0.618	0.640	0.617	0.624
N	831	831	699	699	801	801

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.12: The effects of treatments on unanimity by score difference

	Dependent variable: Unanimous decision					
	b/se	b/se	b/se	b/se	b/se	b/se
Org. Values	0.037	0.051	-0.086	-0.096	0.278***	0.279***
	(0.08)	(0.08)	(0.11)	(0.11)	(0.10)	(0.10)
Social Image	0.014	0.026	0.076	0.071	0.242**	0.235**
	(0.07)	(0.08)	(0.11)	(0.11)	(0.09)	(0.09)
Org. Values &	0.018	0.032	0.060	0.052	0.191*	0.195**
Social Image	(0.08)	(0.08)	(0.10)	(0.10)	(0.10)	(0.10)
Controls	No	Yes	No	Yes	No	Yes
Control mean	0.254	0.243	0.362	0.368	0.287	0.287
N	335	335	225	225	263	263

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.7.D Additional follow up results table

Table 4.13: Long-term outcomes and competition performance

	Transition into Self-employment	Self-employed	Wage-employed	Total income (ETB 1000)	Total investment	Total steps Taken	Has Loan
	b/se	b/se	b/se	b/se	b/se	b/se	b/se
Above average control	0.125	-0.050	0.098	20.625	-0.771	-0.522	-0.011
	(0.09)	(0.12)	(0.10)	(12.69)	(1.49)	(1.36)	(0.03)
Above average org. values	-0.050	0.014	0.068	-6.152	0.800	-0.201	-0.012
	(0.09)	(0.12)	(0.09)	(15.23)	(1.51)	(1.34)	(0.04)
Above average social image	-0.118	-0.148	0.086	7.806	-2.401	-0.430	-0.037
	(0.09)	(0.12)	(0.10)	(12.44)	(1.49)	(1.45)	(0.04)
Above average both	0.006	0.112	-0.210**	9.627	0.868	0.711	0.067
	(0.10)	(0.12)	(0.10)	(14.42)	(1.52)	(1.47)	(0.06)
Winner	0.027	0.190	0.095	34.931	3.354	0.963	0.091
	(0.13)	(0.21)	(0.15)	(65.06)	(2.69)	(2.25)	(0.13)
Control mean	0.151	0.389	0.724	23.504	6.060	6.378	1.035
N	89	89	89	89	89	89	89

Notes This table describes the relationship between labor market outcomes and candidate performance in each of the treatment arms. The transition into self-employment is a dummy equal to one if respondents was not in self-employment before but is after the competition. Self-employed and wage-employed are dummies. Total income is monthly income out of any source. Total investment is the amount the respondent has invested in their business since the competition, this is implemented as a Poisson regression to reduce the influence of outliers. Total steps taken is the number of steps (from a fixed list) a respondent has taken starting a business. Has loan is an indicator for whether the respondent has a business loan.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.7.E Details experimental design

4.7.E.1 Protocol for the control group

This subsection details the protocol for the business plan competition for those judges not assigned to a treatment, (the control group). Following this we then detail how the protocol is different for those receiving the social image treatment. There are no further details to report on the organisational values treatment.

Judges are invited in groups of 16 to a hotel in central Addis Ababa. Once the judges arrive at the hotel and register, the assessments will start at a fixed time after which no new judges are included in the main sample. To start, the judges enter a room set up in a classroom style format and watch two videos. The first video explains the full format of the assessments and how the winner will be decided; the second video features a prominent Ethiopian businessman discussing the importance of access to start-up capital for aspiring entrepreneurs in Ethiopia (and, thus, the importance of the business plan competition). After watching these two videos, and after having had an opportunity to ask questions after the first video, the responses to which are strictly pre-defined, the assessments start.

The respondents then start the assessments in which they will watch the three-minute recordings of twelve pairs of candidates in the business plan competition. After watching each pair, which consists of one female and one male candidate, they are asked which of these candidates they want to cast their vote for the business plan competition and how difficult it was to make this decision. The respondent answers these questions without showing the enumerator their answer.

These assessments provide the data for the main experiment, with as primary outcome whether the judge votes for the female candidate. Following the assessments, we then return to one of the pair of candidates and ask the judge to provide feedback. The enumerator observes, using the survey software, who the judge voted for for one pair of candidates. The respondent watches the recordings of these two candidates again, and is asked to write down reasons for their decision and to provide some feedback for the

candidates both on some fixed dimensions and open-ended feedback. Finally, the judge is asked the following question for both a male and female judge:

> Please now imagine a [male/female] judge with [x] years of managerial experience working as [position] in [industry] has also assessed these candidates. We would like to ask you whom you think they would have voted for.

The purpose of this question is to elicit beliefs on how other judges would vote. The specific set-up is designed to mirror the committee treatment. These fields are filled based on the set of other judges who are invited to attend the same assessment session, i.e. if judges A and B both attend the same assessment session, judge A may here see these characteristics for judge B. Note that while the position, industry and years of experience are based on the actual characteristics, the gender may be changed to ensure the judge is asked about both a male and female judge.

We then run a "feedback session" with the individual judge in which they first, privately, write down reasons for their decision. They then give feedback on a number of fixed dimensions taken from Fafchamps and Quinn (2015) before being given the opportunity to provide some open-form feedback to the candidate. This feedback will then, after the conclusion of the full experiment, be shared with the candidate.

To finish the assessment day, judges are asked a number of closed form-questions about the experiment, about their relationships with other judges and on the ambivalent sexism inventory Glick and Fiske (1996); see the separately attached document 'post-competition questions' for the full questionnaire. We implement this after the experiment in order to avoid the possibility of experimenter demand effects.

The entire experiment is conducted in Amharic.

4.7.E.2 Further details on the Social Image Treatment

The protocol for judges assigned to the social image treatment differs from that for individual judges as follows, and is explained in a separate explanatory video to that for individual judges:

- The room will be set up boardroom rather than classroom style so judges can see each other, and crucially the judges with whom they are on a committee.
- At the start of the assessment day, judges are asked to introduce themselves by telling other judges their name, industry and their position at their company. The judges are asked not to share any additional information.
- Judges still go through the decisions individually, but know they are deciding together with two other judges in the room. Before they watch the videos for each pair of candidates, they are shown photo CV's of these two judges. This includes, beyond a photo of the judge, their name, industry and position at their company. They are also reminded their decisions may be shared with these judges in the feedback sessions.
- The judges are on a different random triplet of judges for each pairwise assessment.
- After the twelve assessments, they also re-watch one of the pairs of candidates' responses. They are then asked who they think *the other two judges on this triplet* voted for. They again see the exact same set of characteristics as those the individual judges see based on the photo CV, and now additionally know this judges' name and what they look like.
- The three judges then come together and are told who each of them voted for the pair of candidates whose videos they just re-watched. They are asked to each give reasons for their decisions to each other as the judges not receiving the social image treatment judges did individually. They then, together, provide the same feedback as an individual judge to the candidate.
- The judges know that in the case of a split decision, the candidates will get respectively 2 and 1 point, *i.e.* one point for each vote that is cast for them.

To ensure judges on the social image treatment are aware of the section at the end of the experiment in which their decisions are made public to other judges. To do so, the

feedback session is clearly highlighted in the video at the start of the assessments explaining the protocol. In this video, actors playing judges go through the full experiment and are shown discussing their decisions (with the sound muted; the respondents know these are actors). Beyond this, judges are reminded every time they are shown the judges on their triplet that they may have to give reasons for their decisions to these two judges.

4.7.E.3 Details on randomisation

Judges are first invited to participate as a judge in the business plan competition. We aim to over-sample judges due to high expected non-attendance, thus intending to invite 320 judges in total aiming to have 240 actually attend including 180 male and 60 female judges. After inviting the judges, we randomly assign them to a treatment assigning 20 women and 60 men to each treatment. We thus assign 80 judges assigned to each treatment. Once a judge is assigned to one treatment, they will not be swapped across treatments.

To specify how randomisation is done, once we have a sample of judges we will randomly split judges across the treatments using Stata's "cut" command based on a random uniform variable with no duplicates. Using this command, the set of male and female judges are separately assigned to the four treatments resulting in a quarter of the male and female judges in the sample being assigned to each treatment.

We will run five of six sessions for each treatment arm over the course of two weeks, running two sessions each day. Every two days, fours sessions, one for each of the four treatment arms, will be conducted during this time (except on Sunday). This ensures ex-ante attendance for each treatment arm is expected to be the same, and we impose the same constraints for inclusion on individual and committee judges.

We flexibly assign judges to a specific assessment day. Judges are randomly assigned to and invited for a specific assessment day, but if they cannot attend this slot they are offered to attend one of the other four slots for their treatment.

For each assessment day, we aim to include at least ten male and two female judges. To achieve this, we invite eleven male and four or five female judges to each assessment

day. We will conduct the experiment if more than eight judges show up, if fewer than eight judges attend the assessments will be rescheduled.

Note that to do the feedback sessions at the end of the sessions for the committee treatment, we require a multiple of three judges to attend for each judge can be included in a three-member assessment. To maximise the number of judges we can include, we allow for some judges not to do a feedback session for the committee treatment if they cannot do so as part of a three-member committee.

The final element of randomisation is due to the difficulty in predicting how many judges will attend. On the day, each judges is randomly assigned a "judge number" from one to the number of judges attending. These judges numbers have been randomly pre-assigned to be members of a specific set of triplets assessing specific pair of candidates. This method allows us to flexibly deal with attrition in a logistically achievable way.

Finally, we randomly assign candidates to competitions. We first stratify by gender, to then randomly assign five male and five female candidates to each competition. As the candidates are not present at the competition, there are no logistical constraints in this randomisation.

4.7.E.4 Invitation judges

I am calling on behalf of EconInsights, an Ethiopian research company. Your company has participated in several surveys with us as part of a research project of the University of Oxford over the past years. Most recently, you have helped us rank aspiring managers and entrepreneurs based on hypothetical vignettes. As we told you at the time, the individuals who performed best in the vignettes as an entrepreneur have now been invited for a business plan competition.

We would like to draw on your personal expertise as a successful member of the Ethiopian business community, and invite you to be a judge in the business plan competition. In this role, you would be watching videos of young individuals pitching business plan proposals. We would like you to help us decide who should get a 50.000 Birr grant towards their business plan.

We would ask for you to join us for around three to four hours to do these assessments. We would pay you X Birr for your time in addition to covering transport expenses. You would not have to prepare anything for these assessments. If you are willing to participate, we will call you in the near future to schedule a time and place sometime in the coming month for this competition. Would you be willing to participate?

If yes: We would like to introduce you to the other judges using your professional credentials. Our records say these are as follows:

- Name [X]
- Company [X]
- Position [X]
- Years of experience as a manager [X]

Is this the correct (q1) name, (q2) company, (q3) position and (q4) years of experience? [Enumerator: Update this in our records with follow-up question if not.]

4.7.F Invitation candidates

Dear X,

I am calling on behalf of EconInsight. As you will remember, you were invited to participate in a management challenge where you got to see different scenarios and we recorded you responding to these scenarios on X date. As promised, we showed these recordings to human resources managers of different firms to determine who would participate in a business plan competition designed to support promising young Ethiopian entrepreneurs. Based on the assessment we obtained from HR managers, we're happy to announce that you have been selected to present your business ideas for the business plan competition. We would like to congratulate you, as your good performance means that you can potentially obtain a reward of 50,000 Birr if your business plan is selected by independent judges. I will now offer you some key details to help you prepare your business plan. In this competition, you will go up against nine other candidates for a

chance to win the 50,000 Birr prize. To participate, we will ask you to come to our studio to record a three minutes business proposal before the 11th of March. Our judges, experienced HR managers, will look at your entry in the following month, and we will aim to inform you of the results by the end of April. If you do not show up to the Studio to record your business plan by March 11th , you will not be able to participate and you lose the opportunity to win the prize .

For the competition, we will ask you to prepare a three-minute pitch for your business, also giving a brief introduction of yourself. This can either be for an existing business, or a plan for a new business. In your entry, you should split your presentation between introducing yourself, your business idea (opportunity), target market, potential competition, operations and cost of business. This should all be done at most in three minutes. Do you have any further questions?

If you are interested, you simply need to agree that we can record your proposal and play them for our judges, primarily human resources managers; we would also use your answers, anonymously, as part of our research on business plan competitions and committees in Ethiopia. We will not share with the judges any of the answers you have given in any previous questionnaires. We would like you to come to the studio as soon as possible. Could you attend on X date? Our Studio is located around Meskel flower road, in a building which hosts Kezira advertising on the 6th floor. [note to self: include other landmarks]. Of course, feel free to get in touch with me if you would like me to give you the precise location of the Studio. Also note that we will fully cover your transport expenses to make the trip to the Studio.

We will give you a follow-up call to remind you your appointment day in the day before you are scheduled to come to the studio. Is this the best phone number to reach you?

4.7.F.1 Judge Beliefs

Immediately after completing the twelve assessments, we elicit judges' beliefs about how other judges have made their decisions. Specifically, for one pair of candidates each judge

assesses, we ask the judge to rewatch the two videos and predict the behavior of two other judges. Judges in the social image treatment arm are asked to predict how the two other judges on that they will have to justify their decision for voted. Judges in the individual treatment arm are asked to predict how they believe two hypothetical judges, one male and one female, have decided. For each of these hypothetical judges, we share their gender, position, and amount of work experience, with these characteristics based on those of actual judges in the experiment. This exercise is not incentivised; it is simply a stated prediction. Since the question differs for social image treatment and non social image treatment judges, we analyze the effect of the organisational values prompt for these two groups separately.

To analyze the data, we run the following regression for the subsamples of judges with and without the social image treatment:

$$\texttt{Predict\_Agree}_j = \alpha + \beta_1 \cdot \texttt{Social\_Image}_j + \varepsilon_{jcp} \tag{4.31}$$

For the sample of judges in the committee treatment, half are on a committee with one female judge. We define the indicator $\texttt{Mixed\_Gender}_j$ to equal one when the two other social image judges are of different genders. We then run the following regression:

$$\texttt{Predict\_Agree}_j = \alpha + \beta_1 \cdot \texttt{Social\_Image}_j + \beta_2 \cdot \texttt{Mixed\_Gender}_j + \tag{4.32}$$

$$\beta_3 \cdot \texttt{Social\_Image}_j \cdot \texttt{Mixed\_Gender}_j + \varepsilon_j \tag{4.33}$$

4.7.F.2 Result 1: Without the information treatment, judges believe male and female judges will vote differently.

For non social image treatment judges, the treatment strongly increases the probability that judges expect a hypothetical male and female judge to behave in the same way. Non social image treatment judges are expected to agree only 62% of the time without the organizational values treatment, which increases to almost 90% with the treatment.

For judges in the social image treatment, there is no *average* increase in the probability

of agreement, as shown in column (2). However, column (3) indicates that without the social image treatment, two male judges agree 95% of the time, compared to 70% agreement between a male and a female judge. With the organizational values treatment, judges now believe a male and female judge are slightly more likely to agree than two male judges.

Table 4.14: The effect of the prompt on the agreement of hypothetical judges

	(1) Sample: Non-committee judges Agree (hypothetical) b/se	(2) Sample: Committee judges Agree b/se	(3) Committee judges Agree b/se
Organisational values	0.273***	-0.001	-0.140*
	(0.08)	(0.06)	(0.08)
Mixed-gender committee			-0.240***
			(0.08)
Organisational values × Mixed-gender committee			0.325***
			(0.12)
Constant	0.622***	0.861***	0.955***
	(0.06)	(0.04)	(0.05)
N	102	129	129

The dependent variable equals one if the judge thinks two (hypothetical) other judges will agree. In column (1) this regression is done for the non-committee judges, who are asked about the behaviour of a hypothetical male and female judge. In columns (2) and (3) this is done for the sample of committee judges who are asked about the other two judges on their committee. This is done for the subsample of male judges only. The mixed-gender committee is an indicator for whether the two other judges on the committee have the same gender.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.7.F.3 Result 2: Treatments result in judges believing male and female judges will vote the same way.

Based on Table 4.14, we see that following the treatment, judges believe there is a 90% chance a male and female judge will agree, compared to an 80% chance that two male judges will agree.

We do not report the full result here but note that this effect is not driven by an increased expectation of coordination on specifically male or female candidate

Exploring Heterogeneity in Management: A Bayesian Approach to Identifying Complementarities

Abstract

This paper introduces a Bayesian hierarchical model to estimate latent management quality across four dimensions of management practices using World Management Survey data. By treating management ability as a latent variable, a mixed ordered logit model is employed to explore correlations among different management practices. The findings indicate significant heterogeneity and varied correlations, suggesting some dimensions act as complements while others are substitutes. Additionally, a two-type model reveals clustering in management practices, highlighting differences across firm types. The empirical approach provides a promising approach for further study of firm management, for example by linking management practices to productivity data, using panel data, and leveraging exogenous variation to identify causal effects.

5.1 Introduction

Management practices are key to explaining differences in firm productivity and growth across firms and countries. There is clear empirical evidence for both a positive causal (Bloom, Eifert, Mahajan, McKenzie, and Roberts, 2013b; Bruhn, Karlan, and Schoar, 2018; Giorcelli, 2019; Gosnell, List, and Metcalfe, 2020) and correlational (Bloom and Van Reenen, 2007; Rasul and Rogger, 2018) link between management practices and firm outcomes. However, there has been less attention to the choice of the combination of practices that a firm employs, despite its key role in the management literature (Ennen and Richter, 2010; Porter et al., 1996; Teece, Pisano, and Shuen, 1997).

This paper contributes to the ongoing debate between two models of management: management by design (MBD) and management as a technology (MAT). Management by design typically refers to management choices being contingent on an organisation's environment or history. In contrast, management as technology posits that management operates like a capital stock in which organisations invest, and that good practices can be directly copied by other firms with sufficient investment (Bloom et al., 2016a). These models represent two illustrative, limiting cases; in reality, they are not mutually exclusive, making it an empirical question as to which framework better describes firm behaviour.

To study the combinations of practices that firms choose to employ, I develop an ordered logit mixture model to estimate firms' latent management quality. This model combines data across multiple dimensions of management practices to produce individual measures of management quality for each dimension. The Bayesian hierarchical structure allows me to estimate the distribution of different combinations of practices employed by firms. By assuming that firms are drawn from a population of similar firms, the model addresses concerns about measurement error. This empirical approach captures the heterogeneity in management practices across firms and tests for clustering of practices through a model extension that incorporates firm types.

The empirical analysis in this paper is based on data on manufacturing firms from

the 2006 World Management Survey (WMS), which collects detailed information on management practices across several countries and industries. The survey focuses on four key dimensions of management: monitoring, target setting, lean management, and people management. Monitoring refers to the extent to which firms track performance indicators and utilise them for decision-making. Target setting measures the effectiveness and rigour of performance goals. Lean management captures the practices aimed at streamlining operations, while people management encompasses the processes involved in hiring, developing, and retaining talent. This is an especially interesting setting to explore, given that these practices are perceived as universal best practices across diverse organisational contexts.

I find that the management practices employed by firms are strongly correlated across the four dimensions measured by the World Management Survey. This is particularly pronounced for monitoring practices, which are highly correlated with lean management and target setting. However, people management is less strongly correlated with other practices, suggesting some complementarities between certain management practices. This finding aligns with Hall, Mairesse, and Mohnen (2010). In the model incorporating types, I observe clear heterogeneity across countries in how strongly practices are linked. For instance, in Germany, all practices are highly correlated, whereas in Greece, the correlations between practices are weaker, though monitoring practices remain strongly correlated with lean management and target setting. It is likely that this reflects institutional and cultural differences.

This paper contributes to the debate on management as technology versus management by design through a new estimation approach to study heterogeneity in management practices. Much of the existing literature uses principal component analysis to study heterogeneity in the combinations of management practices within firms and their link to firm observables Bloom et al. (2016a); Bloom and Van Reenen (2007). Bloom et al. (2016a) provides a more in-depth study of the ability of the two frameworks to describe plants using a stylised model of management by design.

The development of a new framework for estimating heterogeneity in the combina-

tions of management practices employed also contributes to the literature on measuring latent traits. The empirical approach sees the measurements in the WMS as proxies for the underlying firm management "quality" in each of these dimensions. The idea of discovering the underlying distribution of a latent variable capturing management quality is similar to the ideas in (Bandiera et al., 2020), who study CEO behaviour and the skills measurement systems literature (Attanasio, Cattan, Fitzsimons, Meghir, and Rubio-Codina, 2015; Cunha, Heckman, and Schennach, 2010). These approaches differ in the exact specification of the empirical models.

The remainder of the paper proceeds as follows. Section 5.2 provides a short review of the literature on complementarities in organisations. Section 5.3 describes the data from the World Management Survey used for this paper. Section 5.4 describes the empirical model. Section 5.5 describes the results from the estimation of this model. Finally, Section 5.6 summarises the findings of this thesis and discusses potential future research that could extend the theoretical and empirical model developed in this chapter.

5.2 Complementarities and management by design

In this section, I review the literature on management by design to further motivate the main empirical analysis in this paper. The literature on management by design considers how firms choose a bundle of management practices—a set of practices and processes to run an organisation—in response to their history and their internal and external environments. The basic premise throughout these models is that there are complementarities between organisational design choices (Brynjolfsson and Milgrom, 2013). These complementarities help explain patterns in organisational design choices and how they fit with the firm's broader environment. This literature predicts that practices tend to cluster—adopting one of two complementary practices increases the likelihood of adopting the other. This framework also allows for practices being substitutes, with one practice making adopting a second practice *less* valuable – for example a firm which can perfectly monitor employees might not find it valuable to also have a strong incentive structure. Understanding these processes across firms is crucial for understanding both the diffusion

of best practices and the portability of management interventions across contexts.

This framework is prominent in the management literature, where terms like "fit" (Porter et al., 1996), "coherence" (Teece et al., 1997), and "complementarities" (Ennen and Richter, 2010) are used in the context of organisational design and the choice of managerial practices. In this literature, these high-dimensional concepts are key to explaining persistent performance differences across organisations, as the precise configurations are difficult to replicate (see also (Rivkin, 2000) for a theoretical explanation). The empirical literature in this area is primarily based on case studies.

The management by design framework does not lend itself straightforwardly to econometric analysis due to the dimensionality of the problem. To capture this complex relationship between management, firm performance, and productivity differences, the economics management by design literature employs a variety of approaches.

Empirically, early papers use case studies to document such complementarities, such as the organisational structure of Lincoln Electric Company (Milgrom and Roberts, 1995) and the post-war Japanese economic organisation (Milgrom and Roberts, 1994). Ichniowski, Shaw, and Prennushi (1997) introduced the 'insider econometrics' methodology to study these complementarities through a detailed investigation of a narrowly defined industry, finding clear evidence for complementarities in steel-processing plants. More recently, Juhász, Squicciarini, and Voigtländer (2024) has provided further evidence in this area.

A large number of papers have been written on human resource practices, IT, and organisational change, focusing on individual organisations (See, for example, Barley (1986), Autor, Levy, and Murnane (2002), Milgrom and Roberts (1995), Siggelkow (2001), Siggelkow (2002b) and Siggelkow (2002a)). These papers find that clusters in organisational practices arise due to the interaction of different practices and characteristics. For example, Autor et al. (2002) examined the different effects of improved computer-based technology on two floors in one bank. On the ground floor, simple, routine tasks of a bank teller were automated or split into several narrow jobs that could not be done by a computer, whereas on the top floor, various management tasks were combined to cre-

ate jobs of greater complexity. This shows how new technological innovations interact with existing systems to result in radically different changes in management practices, even within the same company. These interaction effects are expected to lead to clusters of practices. This literature provides evidence for the need to model the managerial optimisation problem using a complex landscape to capture these interactions.

The literature based on the World Management Survey also identifies suggestive evidence for the management by design perspective.[63] First, using PCA, Bloom, Lemos, Sadun, Scur, and Reenen (2014) find one principal component that loads on all practices ("good" management) and one that loads on monitoring and targets, and negatively on incentives. This can be partially explained by the country of origin of the firm and suggests there are some complementarities across practices or differences in constraints faced by firms. Bloom et al. (2016a) find, using US data, that firms with more fixed capital tend to specialise in targets and monitoring, whereas human capital-intensive sectors focus more on people and incentives management. Lemos and Scur (2018) study management in family firms, showing that these firms, particularly those with a family CEO, are worse managed.[64]

There is also a broad recent literature studying complementarities in organisations. Giorcelli (2019) and Bianchi and Giorcelli (2022) find complementarities between management practices and capital investment, as well as between management practices themselves, using historical data. Brynjolfsson, Aral, and Wu (2010) and Brynjolfsson, Rock, and Syverson (2017) find evidence of complementarities between investment in IT and organisational practices. Garicano and Heaton (2010) support the existence of complementarities between organisational practices and IT usage, using data on a panel of police departments in the US.

In a related strand of research, three papers provide large-sample evidence for the existence of management "styles" driven by the role of individual managers including the second chapter of this thesis as well as Bertrand and Schoar (2003) and Bandiera et al.

[63] Although (Bloom et al., 2016a) identified "management as technology" as a better perspective compared to their definition of management by design.

[64] Although they conjecture that this is likely due to worse people management practices related to reputational concerns, they do not find direct evidence for this.

(2020). Bertrand and Schoar studied management styles by focusing on the manager fixed effect of CEOs on investment, financial, and organisational practices, tracking individual managers over their careers. The paper identifies specific management styles that are correlated with manager fixed effects in performance. This shows that, despite operating in similar environments for the same firm and controlling for several covariates, different managers have a clear effect on organisational practices. The latter paper further develops this through a machine-learning-based manager shadowing exercise, collecting a large dataset on managerial time use through daily phone calls to firms. The authors find evidence for the existence of two types of managers and differences in demand amongst firms for these two types of CEOs.

This literature provides extensive evidence for the existence of complementarities between management practices and organisational structure, but there is a gap in this literature studying these questions using the large-scale dataset collected as part of the World Management Survey. Given the high dimensionality of management practices, more sophisticated estimation approaches are necessary to capture the complex relationships between management practices and firm observables.

5.3 Data

I use the WMS data on manufacturing firms from the 2006 survey wave.[65] This survey provides data on 18 management practices for 3064 firms from 10 countries in Europe and North America. For a detailed explanation of the data collection process and the dataset, see Bloom and Van Reenen (2007) and Bloom et al. (2016a). Here, I give a brief overview of the methodology, dataset and some results. I consider only data on manufacturing firms in a single wave of the World Management Survey to focus on different management styles amongst comparable firms at the same point in time.

The WMS methodology addresses small-sample issues in the previous literature on management practices by conducting a large-scale survey on comparable management

[65] In practice, the observational units of this dataset are plants. As is common in the literature using the World Management Dataset I refer to these as firms throughout.

practices. The answers are recorded based on open-ended questions related to 18 management practices, which are then coded into scores from one to five by enumerators. To ensure the scores are consistent, the interviews are conducted with a second, silent interviewer also scoring the plant based on the interview.

For the manufacturing sector, the WMS questions relate to 18 practices grouped into four key dimensions (Bloom and Van Reenen, 2007):

> First, *Monitoring*: how well do companies track what goes on inside their firms, and use this for continuous improvement? Second, *Target setting*: do companies set the right targets, track outcomes, and take appropriate action if the two are inconsistent? Third, *Incentives/people management*: are companies promoting and rewarding employees based on performance, and systematically trying to hire and retain their best employees?

The fourth dimension is operations. This measures the extent to which firms have implemented *lean* management practices. Table 5.1 lists the 18 topics, groups them by the four dimensions and provides some summary statistics. Appendix Table 5.7 provides further detail on the questions asked.

The Appendix Tables also describe the distribution of plants across countries, ownership types and industries. Appendix Table 5.8 shows that plants are reasonably distributed across 10 countries, including France, Germany, Great Britain, Greece, Italy, Northern Ireland, Poland, Portugal and Sweden. Appendix Table 5.10 shows the distribution of industries the plants are in. Table 5.9 shows that most firms have dispersed shareholders or are family owned, with a broad variation of different ownership arrangements.

Table 5.1: World Management Survey data description

Topic	N	Mean	Std. Dev.
Lean Management			
Introducing Lean (Modern) Techniques	3,056	2.90	1.14
Rationale for Introducing Lean (Modern) Techniques	2,869	3.03	1.20
Process Documentation and Continuous Improvement	3,061	3.27	1.06
Rationale for Introducing Lean (Modern) Techniques	2,869	3.03	1.20
Monitoring			
Performance Tracking	3,059	3.52	1.06
Performance Review	3,050	3.46	1.03
Performance Dialogue	3,028	3.31	1.04
Consequence Management	3,035	3.29	1.00
Targets			
Types and Balance of Targets	3,049	3.09	1.17
Interconnection of Targets	3,047	3.12	1.08
Time Horizon of Targets	3,048	3.19	1.18
Target Stretch	3,045	3.13	0.96
Clarity and Comparability of Goals	3,046	2.59	1.04
Incentives and Talent Management			
Instilling a Talent Mindset	3,045	2.49	1.08
Building a High-Performance Culture through Incentives	3,039	2.57	1.10
Removing Poor Performers	3,053	3.23	1.08
Developing Talent and Promoting High-Performers	3,044	3.13	1.02
Distinctive Employee Value Proposition	3,019	3.16	0.97
Retaining Talent	3,038	2.53	1.08

Notes: This table summarises the questions captured by the World Management Survey for manufacturing firms in 2006. Note that just under 1% of the data is missing.

5.4 An empirical model of heterogeneity in management practices

To efficiently aggregate the information content of the World Management Survey to study whether firms choose to employ heterogeneous combinations of practices, I develop a Bayesian hierarchical model to capture the *combination* of management practices chosen by a firm. The assumption of this estimation approach is that a firm's management quality can be summarised by latent management scores for each of the four dimensions of management measured by the World Management Survey. This model allows us to both capture cross-firm variability of management practices, and how individual firms deviate from these general patterns of management practices.

This method provides three main benefits. First, it allows us to learn about the covariance structure of management practices across firms–and whether the quality of practices is strongly connected across all dimensions or whether some dimensions are less correlated. Second, since the World Management Survey (WMS) contains significant measurement error (Bloom and Van Reenen, 2007), the joint model structure helps estimate the management quality for individual dimensions by leveraging shared information across dimensions.[66] Third, the model uses more information in estimating the model by incorporating the ordinal (Likert) structure of the data and using information across multiple dimensions.

Specifically, we observe eighteen measures of management practices $\boldsymbol{y}_f$ for F firms, f, all scores on a Likert scale from one to five. For each firm, these measures are drawn from four dimensions, d, of management. We can thus denote $\boldsymbol{y}_f = [\boldsymbol{y}_f^1, \boldsymbol{y}_f^2, \boldsymbol{y}_f^3, \boldsymbol{y}_f^4]$. These four dimensions are respectively operations, targets, monitoring and incentives. Dimension d has $J(d)$ elements.[67] The model assumes each firm belongs to one of several possible management "types", and within each type, practices follow certain correlated

[66] This is more effective the more correlated the true values of the practices are and the less correlated the measurement errors, as this allows the estimator to "smooth out" the noise rather than propagate it. Appendix Table 5.11 shows the correlation structure of the differences between these repeated measurements across these practices, the average correlation is moderate at 0.24. This also implies any estimator is likely to overestimate the correlation between characteristics.

[67] Specifically, J() equals 3 for operations, 4 for targets, 5 for monitoring and 6 for talent management.

patterns. The generative model relies on the following assumptions:

Assumption 1 *The individual means for the four dimensions for each pure type,* $\boldsymbol{\mu}_f^t = \begin{bmatrix} \mu_f^{t1} & \mu_f^{t2} & \mu_f^{t3} & \mu_f^{t4} \end{bmatrix}$*, follow a multivariate normal distribution denoted as:*

$$\boldsymbol{\mu_f^t} \sim MVN(\mathbf{0}_{4\cdot 1}, \boldsymbol{\Sigma}_{4\cdot 4}^t)$$

Assumption 2 *Within each dimension, d, the observed ordinal outcomes are determined by a set of cut-off points, denoted by* $\boldsymbol{\gamma}^d = \begin{bmatrix} \gamma_1^d & \gamma_2^d & \gamma_3^d & \gamma_4^d \end{bmatrix}$*. These cut-off points partition the underlying latent variable* mu_{td}^f *into ordinal categories.*

Assumption 3 *The characteristics in one dimension, d, share a cut-off:* $\boldsymbol{\gamma}^d = \begin{bmatrix} \gamma_1^d & \gamma_2^d & \gamma_3^d & \gamma_4^d \end{bmatrix}$.

Assumption 4 *Difference in the mean of the observed values across characteristics within a dimension are captured using a linear shift of the latent variable:* $\boldsymbol{\beta}$.

Assumption 5 *Each firm is one of T types of firms. The probability that each firm is each type is determined by their type score* $\boldsymbol{\theta}_f^t$*:*

$$\boldsymbol{\theta}_f^t = \alpha + \boldsymbol{X}_f' \boldsymbol{\delta}^t + \boldsymbol{\epsilon}_f \tag{5.34}$$

where $\boldsymbol{X}$ *is a set of firm observables. This is mapped to a type probability using a softmax transformation. To ensure identification, the vector of parameters is normalised to zero for the first type.*

Assumption 6 *The error term for each firm is drawn from a logistic distribution.*

Based on this, and conditional on type, for each measure in $\boldsymbol{y}_f$ the enumerator observes a latent management score distributed as:

$$y_{fdj}^* = \mu_f^d + \beta_{dj} + \epsilon_{fdj}, \tag{5.35}$$

Where ϵ_{fdj} follows a standard logistic distribution. This is mapped to the observed response variable based on the cut-offs $\boldsymbol{\gamma^d}$:

$$y_{fdj} = \begin{cases} 1 & \text{if } y^*_{fdj} \leq \gamma^d_1 \\ 2 & \text{if } y^*_{fdj} \in (\gamma^d_1, \gamma^d_2] \\ 3 & \text{if } y^*_{fdj} \in (\gamma^d_2, \gamma^d_3] \\ 4 & \text{if } y^*_{fdj} \in (\gamma^d_3, \gamma^d_4] \\ 5 & \text{if } y^*_{fdj} > \gamma^d_4 \end{cases} \tag{5.36}$$

The probability of observing a specific management practice outcome for firm f, dimension d, and practice j is given by:

$$P(y_{fdj}|\mu^{td}_f, \boldsymbol{\gamma}^d, \beta_{dj}) = \Lambda\left(\gamma^d_{y_{fdj}} - \beta_{dj} - \mu^{td}_f\right) - \Lambda\left(\gamma^d_{y_{fdj}-1} - \beta_{dj} - \mu^{td}_f\right), \tag{5.37}$$

where $\gamma^d_5 = \infty$ and $\gamma^d_0 = -\infty$ and $\Lambda(\cdot)$ is the logistic cumulative distribution function (CDF).

The probability that firm f is of type t given its observables $\mathbf{X}_f$ is:

$$\pi_{ft} = P(T_f = t|\mathbf{X}_f) = \frac{\exp\left(\alpha + \mathbf{X}^\top_f \delta^t\right)}{\sum_{t'=1}^{T} \exp\left(\alpha + \mathbf{X}^\top_f \delta^{t'}\right)}, \tag{5.38}$$

where for identification, parameters α and $\boldsymbol{\delta}$ for the first type ($t = 1$) are set to zero.

The likelihood of observing all responses for firm f is the weighted sum over all types:

$$P(\mathbf{y}_f|\gamma, \beta, \delta, \alpha, \mathbf{X}_f) = \sum_{t=1}^{T} \pi_{ft} \prod_{d=1}^{4} \prod_{j=1}^{J_d} P(y_{fdj}|\mu^{td}_f, \gamma^d, \beta_{dj}),$$

This yields the total probability of observing the dataset conditional on the parameters:

$$P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{f=1}^{F} \sum_{t=1}^{T} \prod_{d=1}^{4} \prod_{j=1}^{J_d} P(y_{fdj}|\boldsymbol{\mu^t_f}, \boldsymbol{\gamma}, \boldsymbol{\beta}), \tag{5.39}$$

Taking the logarithm of the product of these probabilities across all firms yields the

total log-likelihood:

$$LL = \sum_{f=1}^{F} \ln \left(\sum_{t=1}^{T} \pi_{ft} \prod_{d=1}^{4} \prod_{j=1}^{J_d} \left[\Lambda \left(\gamma^d_{y_{fdj}} - \beta_{dj} - \mu^{td}_f \right) - \Lambda \left(\gamma^d_{y_{fdj}-1} - \beta_{dj} - \mu^{td}_f \right) \right] \right), \quad (5.40)$$

This function incorporates the probability that each firm is of a certain type based on its characteristics, π_{ft}, the likelihood of observing each response given the latent management score, thresholds, and item-specific shifts. It then sums over all types, weighting by their probability. This function is then used as part of a Bayesian estimation framework, incorporating prior distributions for the model parameters. In this setup, since the latent variables μ^{td}_f are treated as parameters within the Bayesian model, we directly sample them from their posterior distributions during the inference process. Therefore, there's no need to integrate over their distributions explicitly because the uncertainty in μ^{td}_f is inherently accounted for through the joint posterior sampling of all model parameters.

I use relatively standard prior distributions for the parameters.[68] For the hyperprior for the variance matrix of the individual means, I decompose the variance matrix into a correlation and scale matrix, and place a uniform LKJ-prior on the correlation matrix and normalise the scale matrix to one for identification.[69] For the other parameters I use a normal distribution with variance 1; centred at zero, except for the cut-points where these are centred around the quartiles of the logistic distribution.

The model is estimated using Hamiltonian Monte Carlo (HMC) via the Stan probabilistic programming language. Stan employs the No-U-Turn Sampler (NUTS), an adaptive variant of HMC, to efficiently generate samples from the posterior distribution of the parameters (Hoffman, Gelman, et al., 2014). Appendix 5.7.D provides additional details on the exact constraints and reparameterisations used in the estimation process. Appendix 5.7.E provides further details on diagnostic checks and convergence analysis.

There are a number of key challenges in estimating this model that should be noted. The first is that jointly estimating the type probabilities and type characteristics

[68] This setup is similar to Meager (2019)

[69] Without this normalisation, the cut-offs $\boldsymbol{\gamma}^d$ and the variance are not jointly identified.

leads to slower exploration of the parameter space which is resolved by increasing the number of samples taken with the HMC sampler.[70] The second issue is that the resulting distribution is multimodel, as switching the type labels does not alter the log-likelihood. I deal with this by first asserting that no label switching takes place, and then re-labelling the chains appropriately.

5.4.1 Linking model parameters with the alternate hypothesis

To aid in the interpretation of the model, I consider a few hypothetical correlation matrices and what these would tell us about the world. These three matrices are reported in Table 5.2.

[70] Throughout the estimation, I run a longer chain and keep every tenth iteration to decrease autocorrelation without increasing the size of the resulting chain, and thus the size of the resulting file.

Table 5.2: Hypothetical structures of the correlation between dimensions

	Independence between dimensions			
	Lean	Targets	Monitoring	Incentives
Lean	1.00			
Targets	0	1.00		
Monitoring	0	0	1.00	
Incentives	0	0	0	1.00

	The management as technology structure			
	Lean	Targets	Monitoring	Incentives
Lean	1.00			
Targets	ρ	1.00		
Monitoring	ρ	ρ	1.00	
Incentives	ρ	ρ	ρ	1.00

	The management by design structure			
	Lean	Targets	Monitoring	Incentives
Lean	1.00			
Targets	0.8	1.00		
Monitoring	0.8	0.2	1.00	
Incentives	0.8	0.5	0.5	1.00

Notes This table describes the three hypothetical correlation matrices for the one-type model described in Section 5.4.

First, consider the case with no correlation between traits across dimensions. In this scenario, firms make independent decisions regarding each management practice dimension. The adoption of practices in operations, targets, monitoring, and incentives occurs without influence from the other dimensions, implying a modular approach to

management where investment in one area does not imply investment in others.

Next, consider the case where the correlation between dimensions is equal to ρ. This uniform correlation supports the MAT hypothesis: firms exhibit a consistent level of investment across all dimensions, reflecting a unidimensional management quality. In this context, enhancing practices in one area is inherently linked to improvements in others, perhaps due to decreasing marginal returns or increasing marginal costs of investing in each individual practice.

Finally, the management by design structure presents a more nuanced correlation matrix. The high correlations (e.g., 0.8) between Lean (operations) management and other dimensions indicate that firms prioritising operational excellence tend to complement this with strong practices in targets and incentives. However, the lower correlations between targets and monitoring suggest these practices may act as substitutes. The latter might happen as strong monitoring reduces the need to motivate employees with specific targets. All correlations are still positive to reflect that a better-managed firm is likely to invest more in all dimensions of management.

5.5 Results on heterogeneity in management practices

In this section, I first discuss the results from the estimation of a single-type version of this model, with some discussion of differences in patterns of management practices across different types of firms. Next, I turn to a two-type model of firms to further study these relationships.

Turning to the single-type version of the empirical model, I first discuss the estimated correlation matrix between the quality of management practices across directions. I then study how the management practices employed by different plants in terms of both their observable characteristics and location deviate from this average distribution.

Table 5.3 details the estimated covariance matrix Σ. This implies that the correlation implies that in monitoring is particularly correlated with other dimension including lean management and target setting, but that lean and incentives management are significantly less correlated with other dimensions of management.

Table 5.3: Estimated correlation between latent means

	Lean	Targets	Monitoring	Incentives
Lean	1.00			
Targets	0.73	1.00		
Monitoring	0.93	0.91	1.00	
Incentives	0.54	0.70	0.63	1.00

Notes This table describes the estimated correlation matrix for the model described in Section 5.4 with $T = 1$. The reported correlation is calculated individually for each draw based on the Cholesky factor estimated, and then averaged.

This result implies that monitoring is highly correlated with lean management and target-setting, lean management and target setting are also reasonably highly correlated, and practices related to incentives are far less related to the quality in the other dimensions. Similarly, Hall et al. (2010) find, using Principal Component Analysis, that the variation in management practices can be explained using one component that loads on all questions and can be interpreted as "good management", and one factor that loads positively on monitoring and targets and negative on incentives, suggesting an element of specialisation.

To test whether the model with different correlations between dimensions is appropriate, or whether the correlation structure can be simplified I do the following. I perform a Likelihood Ratio Test for the null-hypothesis that the restricted model with equal off-diagonal elements (ρ) fits the data just as well as the unrestricted model, with the alternative hypothesis that the unrestricted model fits the data better. I calculate the likelihood ratio which has a chi-squared distribution with five degrees of freedom under the null-hypothesis – the off-diagonal elements of the correlation matrix. The test statistic of 151.7 far exceeds the critical value of 15.1 at a 1% significance level.

5.5.1 Heterogeneity in management quality across firms

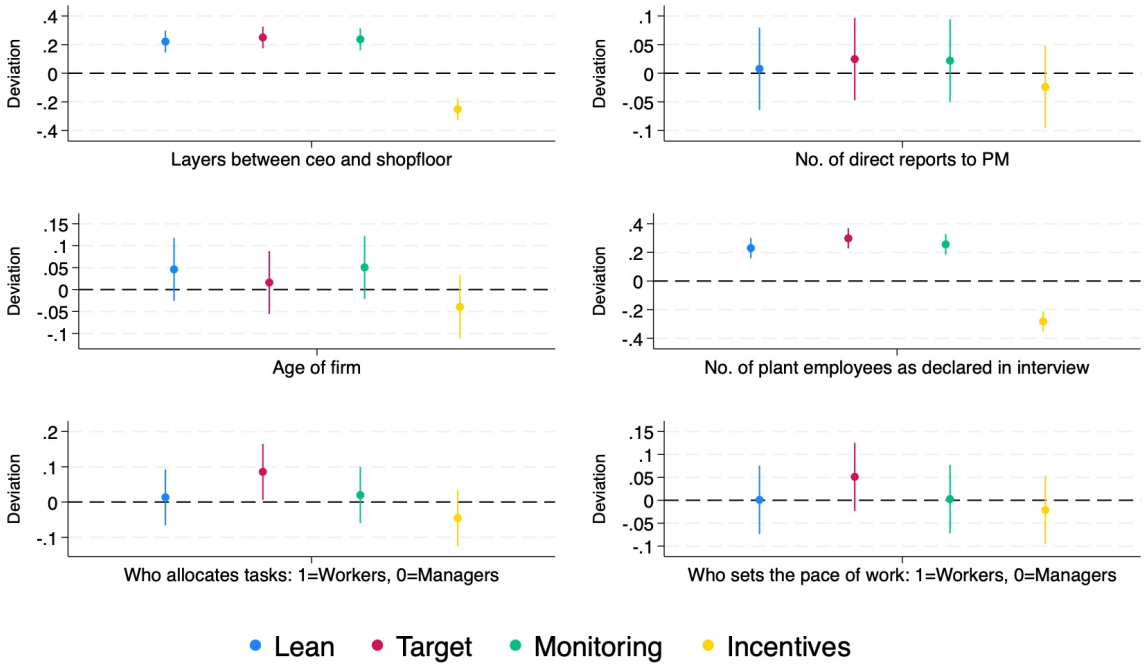
After fitting the model, I conduct the following exercise. Based on the estimated model parameters, I first predict the score for each plant for each dimension of management as a function of their score for the other dimensions. I then calculate the difference between the estimate of the *expected* score and the *actual* score, $\tilde{\boldsymbol{\mu}}_i$. This measure captures how a the distribution of a plant's management quality across dimensions deviates from the average across the full sample of firms. To analyse this data, I run regressions of the following form, regressing a number of observable characteristics X_i on $\tilde{\boldsymbol{\mu}}_i$. I include country ($\gamma^c$) and industry ($\gamma^{\text{ind}}$) fixed effects.

$$\tilde{\mu}_{id} = \alpha_d + \beta X_i + \gamma^{\text{ind}}_{id} + \gamma^c_{id} + \varepsilon_{id} \tag{5.41}$$

Indicating each plant with $i = \{1, ..., N\}$ and each dimension with $d = \{1, ..., D\}$. Each of the variables $X(i)$ is a dummy for whether or not the plant is below the median for a specific plant observable.

Figure 5.1 reports the results of this exercise graphically. Different types of firms appear to particularly choose different combinations of monitoring and incentives practices. Firms with more layers in their hierarchy, more plant employees and more worker autonomy tend to have relatively better monitoring and targets practices, but worse practices related to incentives.

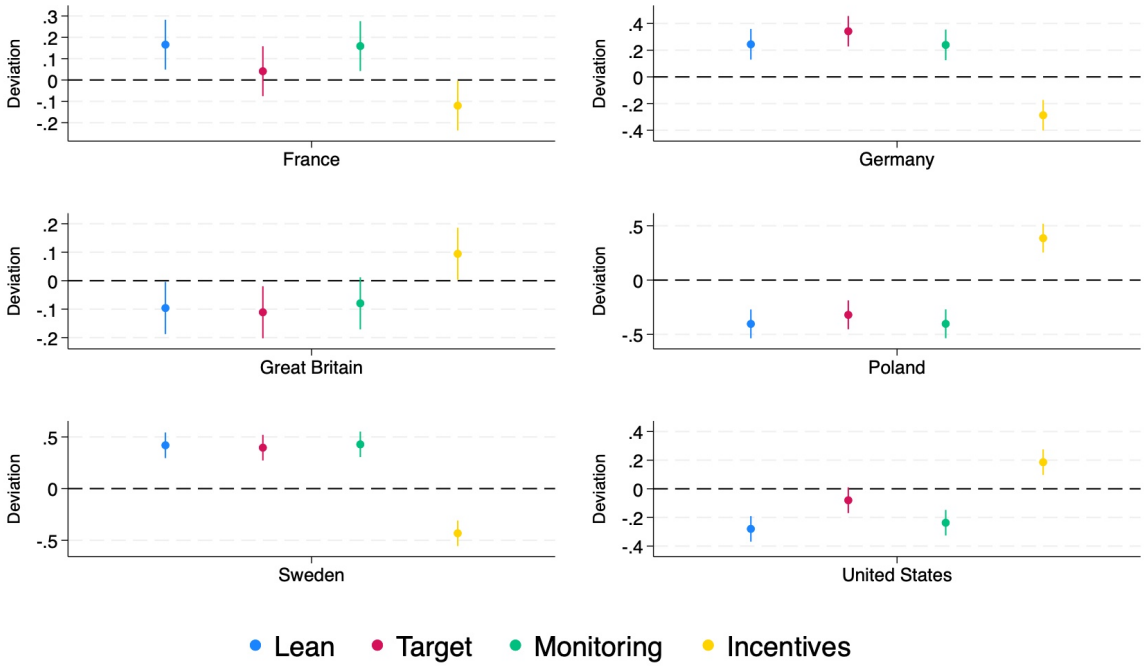
Figure 5.1: Latent management scores and firm observables



Notes This table reports the coefficients of a regression of the WMS score (WMS), the average latent mean across the four dimensions (Overall), and the score for each of the four dimensions. The overall world management score is calculated by normalising the score for each practice to a standard normal distribution, taking the average and normalising this average to a standard normal distribution. The overall score is calculated as the linear mean of the four management scores. Appendix Figures 5.3, 5.4 and 5.5 show these results are robust to the inclusion of respectively only industry dummies, ownership and country fixed effects.

We can perform a similar exercise using country dummies, using a country dummy as our set of covariates X_i and including only industry fixed effects.

Figure 5.2: Standardised management scores by country



Notes This table shows the OLS coefficient of a regression of the deviation from the expectation of the management score in each dimension of management on a country dummy for six countries. This specification includes industry fixed effects. All dependent variables are normalised to have mean zero variance one. Appendix Table 5.6 shows that this result is robust to the inclusion of different sets of fixed effects.

Here, we find clear heterogeneity in the combination of practices that firms in different countries choose to employ. Again, a lot of the heterogeneity is driven by differences in the quality of the incentives practices across countries, with Sweden, Germany and France having relatively "poor" incentives practices, and Great Britain, Poland and the United States having relatively "good" practices. The deviations from the mean of the three other characteristics are more heterogeneous across countries. This is likely related to the finding in Bloom, Genakos, Sadun, and Van Reenen (2012b) that light labour market regulation is related to high scores on incentives, but not monitoring or target management.

5.5.2 An MAT and an MBD type

To further explore this heterogeneity in management practices across countries, I estimate a two-type extension of the model using a set of dummies for the country in which the firm is located in the vector $\boldsymbol{X}$. This version of the model allows the covariance structure between practices to explicitly differ across countries, by allowing each country to be a combination of two pure types.[71]

The estimation returns two clearly distinct types of correlation structures for the two types of firms, described in Table 5.4. I interpret these two types as a management by design type, in which practices are loosely linked, and a management as technology type, for which all management practices move closely together. For the management by design type only monitoring is fairly strongly correlated with lean and target practices. In line with the results in Figure 5.1, we find that firms in Poland and the US versus Germany and France are relatively distinct in the probability they have of being drawn from the MAT type.

Table 5.4: Estimated correlation structure for the two types

	'Management by Design' type				'Management as Technology' type			
	Lean	Targets	Mon.	Incentives	Lean	Targets	Mon.	Incentives
Lean	1				1			
Targets	0.22	1			0.94	1		
Monitoring	0.77	0.71	1		0.97	0.98	1	
Incentives	0.12	0.32	0.19	1	0.74	0.86	0.83	1

Notes: This table describes the estimated correlation matrices for the two types model described in Section 5.4 for $T = 2$. The vector $\boldsymbol{X}$ contains country dummies, excluding the dummy for the United States for identification.

Table 5.5 shows the weights on the MAT type for the countries included in the sample. I find that although all countries have a weight of at least 50% on the MAT

[71] Appendix 5.7.E details convergence of the model, the model exhibits high autocorrelation as is common for mixture models but appears to converge well and shows no signs of label switching.

type, this weight ranges from 50% for Poland to 88% for Germany.

Table 5.5: Weights on the MAT type by country

Country	Weight on MBD Type	Weight on MAT Type
France	0.23	0.77
Germany	0.12	0.88
Great Britain	0.24	0.76
Greece	0.26	0.74
Italy	0.43	0.57
Northern Ireland	0.30	0.70
Poland	0.50	0.50
Portugal	0.35	0.65
Sweden	0.38	0.62
United States	0.38	0.62

Notes: This table reports the estimated probability that a firm in each country is estimated to be in each type. This is calculated by determining the type score for each firm before taking a softmax transformation.

To validate these findings, I compare the two countries closest to these extremes. Table 5.6 reports the correlation in the average management score across dimensions taken directly using the ordinal data for these two countries. In these raw correlations, we observe a comparable correlation of monitoring with targets and operations for the two countries. The correlations for between the remaining practices are, as predicted by the model, lower for Poland. These results mirror the findings in Figure 5.2 that German and Polish firms are managed quite differently. Similar patterns hold for the other countries as a function the probability of being each type. Appendix Table 5.14 reports these correlations for the full list of countries.

Table 5.6: Correlation between management scores for German and Polish firms

	Germany				Poland			
	Lean	Targets	Mon.	Incent.	Lean	Targets	Mon.	Incent.
Lean	1				1			
Targets	0.55	1			0.4	1		
Monitoring	0.62	0.65	1		0.6	0.55	1	
Incentives	0.53	0.60	0.58	1	0.36	0.53	0.53	1

Notes This table describes the correlation matrices between raw management scores for German and Polish firms.

5.6 Discussion

In this chapter, I introduce a Bayesian hierarchical model to combine measurements of a range of management practices into a correlated set of management scores, representing distinct dimensions of management practices. This methodology enables efficient aggregation of individual practices, addressing the challenge of non-cardinality in management metrics that previous approaches have not incorporated. By treating management ability as a latent variable, I employed a mixed ordered logit model to estimate the distribution of this latent variable. This empirical strategy, designed to counter potential measurement issues, enables testing of predictions in the literature on complementarities regarding the combinations of management practices employed by firms.

The results from the one-type model provide a number of insights into the correlations between different dimensions of management practices. Specifically, monitoring practices are found to be strongly correlated with both target setting and lean management, suggesting that plants often bundle these practices together to achieve comprehensive performance tracking and goal alignment. In contrast, practices related to incentives show weaker correlations with other dimensions, indicating that plants may view incentives as less directly integrated with other management strategies. This variation in correlations highlights the presence of specific complementarities, with some practices

reinforcing each other while others remain more independently adopted. These findings suggest that while plants aim for a cohesive management structure, they selectively emphasise certain dimensions over others, reflecting strategic differences in their approach to management. These results are in line with the existing literature on this topic.

The use of a two-type model offers preliminary evidence of clustering in management practices. This exploratory analysis reveals significant differences in the variance and the combinations of practices across plants by their country of origin.

Future research is needed to unpack the sources of this heterogeneity, particularly how different constraints influence the combinations of management practices that firms employ. One implication of the varying correlations between management dimensions is that either (i) the cost of improving practices differs across firms, or (ii) complementarities between management dimensions lead firms to bundle specific practices together. While this chapter cannot definitively distinguish between these possibilities, a potential avenue for future research would involve examining the relationship between estimated latent management quality and firm productivity to identify complementarities. This would provide further insights into whether the heterogeneity observed in management practices reflects optimal responses to firm-specific constraints or intrinsic complementarities amongst practices.

Specifically, three extensions of the analysis appear valuable. First, this analysis does not use the panel dimension of the World Management Survey. This dimension allows for an analysis of changes in the combination of practices employed by types of firms over time, and which firms change their type over time. Second, linking these data to productivity data would allow for studying whether alignment with the average distribution of practices increases productivity, and whether specific types of firms benefit from adopting a different combination of practices than the average firm. Finally, the work in this paper is entirely descriptive; finding causal evidence through IV approaches or policy changes would help understand what drives heterogeneity in the management practices employed by firms.

5.7 Appendices

5.7.A World Management Survey questions

Table 5.7: World Management Survey questions by topic

Topic	Detailed description
Lean Management	
Introducing Lean (Modern) Techniques	*Tests how well lean manufacturing techniques have been introduced.*
Rationale for Introducing Lean (Modern) Techniques	*Tests the motivation and impetus behind changes to operations.*
Process Documentation and Continuous Improvement	*Tests processes for and attitudes to continuous improvement and documentation.*
Monitoring	
Performance Tracking	*Tests whether performance is tracked using meaningful metrics and with appropriate regularity.*
Performance Review	*Tests whether performance is reviewed with appropriate frequency and communicated to staff.*
Performance Dialogue	*Tests the quality of review conversations.*
Consequence Management	*Tests whether differing levels of performance lead to different consequences.*
Targets	
Types and Balance of Targets	*Tests whether targets cover a broad set of metrics and are balanced between financial and non-financial goals.*
Interconnection of Targets	*Tests whether targets are tied to the organisation's objectives and cascade down the organization.*
Time Horizon of Targets	*Tests whether the firm has a '3 horizons' approach to planning and targets.*
Target Stretch	*Tests whether targets are appropriately difficult to achieve and based on a solid rationale.*
Clarity and Comparability of Goals	*Tests how understandable performance measures are and whether performance is communicated openly.*
Incentives and Talent Management	
Instilling a Talent Mindset	*Tests what emphasis is placed on overall talent management within the organisation.*
Building a High-Performance Culture through Incentives	*Tests whether there is a systematic approach to identifying good and bad performers and rewarding them.*
Removing Poor Performers	*Tests how well the organisation is able to deal with underperformers.*
Developing Talent and Promoting High-Performers	*Tests whether promotion is performance-based and whether talent is developed within the organization.*
Distinctive Employee Value Proposition	*Tests the strength of the employee value proposition.*
Retaining Talent	*Tests whether the organization will go out of its way to keep its top talent.*

5.7.B Description of plants

Table 5.8: Number of plants by country

Country in which plant is located	**No.**
France	323
Germany	335
Great Britain	623
Greece	187
Italy	200
Northern Ireland	17
Poland	238
Portugal	175
Sweden	285
United States	681
Total	3,064

Notes: This table shows the distribution of plants by country in the dataset.

Table 5.9: Ownership structure of firms

Firm ownership type	**No.**
Dispersed shareholders	1,078
Family owned, family CEO	272
Family owned, primogeniture CEO	231
Founder owned, founder CEO	304
Private equity/venture capital	186
Private individuals	485
Other (all categories < 5%)	508
Total	3,064

Notes: This table shows the distribution of plants by ownership type in the dataset. The "Other" category was created by grouping ownership categories that each comprised less than 5% of the total sample size (i.e., fewer than 150 observations out of 3,064). The following ownership groups were included in "Other": Banks/holdings/financial institutions (41), Employees/coop (72), Family owned, external CEO (97), Foundation/research institute (14), Founder owned, external CEO (26), Founder owned, primogeniture CEO (2), Government (66), Joint venture (43), Managers (112), Other (35).

Table 5.10: Industry distribution by aggregated two-digit SIC codes

SIC Code	Industry	Count	Percentage
20	Food and kindred products	348	11.36 %
22	Textile mill products	88	2.87 %
23	Apparel and other finished products	67	2.19 %
24	Lumber and wood products	84	2.74 %
25	Furniture and fixtures	66	2.15 %
26	Paper and allied products	158	5.16 %
27	Printing and publishing	114	3.72 %
28	Chemicals and allied products	323	10.54 %
30	Rubber and miscellaneous plastics products	202	6.59 %
32	Stone, clay, glass, and concrete products	127	4.14 %
33	Primary metal industries	130	4.24 %
34	Fabricated metal products	293	9.56 %
35	Industrial and commercial machinery	347	11.33 %
36	Electronic and other electrical equipment	257	8.39 %
37	Transportation equipment	187	6.10 %
38	Instruments and related products	146	4.77 %
39	Miscellaneous manufacturing industries	100	3.26 %
Other	Other (all codes $< 1\%$)	27	0.88 %

Notes This table describes the set of manufacturing firms across which plants are distributed. The other category contains 27 plants.

5.7.C Correlations across practices between differences in scores across repeat interviews

Table 5.11: Correlation in measurement errors based on repeat interviews

	Lean			Monitoring				Targets					Incentives					
	1	2	3	1	2	3	4	1	2	3	4	5	1	2	3	4	5	6
Lean 1	1.00																	
Lean 2	0.39	1.00																
Lean 3	0.37	0.36	1.00															
Monitoring 1	0.19	0.16	0.35	1.00														
Monitoring 2	0.28	0.14	0.32	0.42	1.00													
Monitoring 3	0.30	0.20	0.39	0.39	0.57	1.00												
Monitoring 4	0.24	0.22	0.30	0.27	0.40	0.45	1.00											
Targets 1	0.18	0.15	0.30	0.26	0.28	0.30	0.27	1.00										
Targets 2	0.18	0.18	0.28	0.26	0.34	0.34	0.23	0.45	1.00									
Targets 3	0.12	0.21	0.29	0.18	0.31	0.33	0.24	0.33	0.37	1.00								
Targets 4	0.21	0.10	0.18	0.14	0.16	0.14	0.33	0.27	0.31	0.25	1.00							
Targets 5	0.19	0.27	0.12	0.14	0.19	0.21	0.32	0.19	0.26	0.34	0.27	1.00						
Incentives 1	0.26	0.17	0.25	0.14	0.30	0.26	0.30	0.26	0.23	0.27	0.16	0.27	1.00					
Incentives 2	0.15	0.10	0.08	0.08	0.25	0.26	0.31	0.16	0.17	0.13	0.06	0.28	0.27	1.00				
Incentives 3	0.17	0.10	0.27	0.29	0.16	0.28	0.40	0.22	0.20	0.10	0.23	0.14	0.12	0.18	1.00			
Incentives 4	0.26	0.24	0.30	0.13	0.27	0.31	0.43	0.33	0.38	0.27	0.25	0.27	0.32	0.25	0.45	1.00		
Incentives 5	0.24	0.18	0.27	0.21	0.21	0.33	0.32	0.18	0.21	0.16	0.20	0.08	0.13	0.19	0.30	0.35	1.00	
Incentives 6	0.25	0.19	0.18	0.09	0.25	0.27	0.24	0.16	0.16	0.15	0.16	0.22	0.16	0.29	0.14	0.29	0.32	1.00

Notes: This table reports the correlations in measurement errors across practices, based on repeat measurements by two different enumerators of the same practice at the same plant with different respondents.

5.7.D Details on the estimation procedure

In this appendix, I describe the constraints imposed on the parameters and the implementation strategies that ensure efficient estimation.

Several constraints are imposed on the parameters to ensure model identification and improve estimation efficiency. I order the cut-off points γ_k^d within each dimension ($\gamma_1^d < \gamma_2^d < \gamma_3^d < \gamma_4^d$) to reflect the ordinal nature of the Likert scale used in the WMS. For item-specific intercepts β_{dj}, I constrain them by setting $\beta_{d1} = 0$ for the first practice ($j = 1$), which serves as a reference point and prevents confounding between the intercepts

and the latent scores. I center the latent management scores $\boldsymbol{\mu}_f^t$ at zero ($\mathbb{E}[\mu_f^{td}] = 0$ for all dimensions d and types t) and fix the scale by setting the marginal variances in the covariance matrices $\boldsymbol{\Sigma}^t$ to one ($\text{diag}(\boldsymbol{\Sigma}^t) = 1$). This approach avoids identification issues between the scale of the latent scores and the cut-off points.

To capture the covariance structure among the management dimensions, I decompose the covariance matrices $\boldsymbol{\Sigma}^t$ into correlation matrices $\boldsymbol{\Omega}^t$, such that $\boldsymbol{\Sigma}^t = \boldsymbol{\Omega}^t$. I place an LKJ prior with a shape parameter $\eta = 1.5$ on the correlation matrices $\boldsymbol{\Omega}^t$, which expresses my prior belief that favors modest correlations while allowing flexibility.

To improve the efficiency of the Hamiltonian Monte Carlo (HMC) sampling and address potential issues related to funnel-shaped posterior distributions, I employ a non-centered parameterisation for the latent management scores. Instead of sampling $\boldsymbol{\mu}_f^t$ directly, I introduce standardised variables $\boldsymbol{\mu}_{\text{raw},f}^t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and obtain the latent scores via $\boldsymbol{\mu}_f^t = \boldsymbol{\mu}_{\text{raw},f}^t \boldsymbol{L}^t$, where $\boldsymbol{L}^t$ is the Cholesky factor of the correlation matrix $\boldsymbol{\Omega}^t$. This reparameterisation decouples the latent scores from the covariance parameters, leading to more efficient sampling and better convergence properties.

I estimate the model using Stan, which implements HMC via the No-U-Turn Sampler (NUTS). I run multiple parallel chains with sufficient iterations to ensure convergence, and I monitor diagnostics such as the Potential Scale Reduction Factor (PSRF) and the Effective Sample Size (NEFF). I apply thinning when necessary to reduce autocorrelation in the chains. In models with multiple types, label switching can occur due to symmetry in the likelihood function. To address this, I check for label switching in the MCMC samples, which does not occur, and relabel chains as necessary before analysis.

I center firm-level covariates $\boldsymbol{X}_f$ to reduce correlation between intercept and slope parameters in the type probability model, improving sampling efficiency. I use relatively diffuse priors for the parameters to reflect prior uncertainty while ensuring identifiability. For the item-specific intercepts β_{dj} (for $j > 1$), I use normal priors centered at zero with moderate variance. Similarly, I use normal priors for the cut-off points γ_k^d, centered around the quartiles of the logistic distribution, and enforce ordering constraints.

5.7.E Convergence

I analyse convergence using the Potential Scale Reduction Factor (PSRF) and Effective Sample Size (NEFF). PSRF, also known as the Gelman-Rubin diagnostic, is used to evaluate whether the chains have converged to the same distribution by comparing the variance between chains with the variance within each chain. The theoretical formula for PSRF is given by:

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}}, \tag{5.42}$$

where W represents the within-chain variance, which captures the variability of the samples within each chain, and $\hat{V}$ is the total variance estimate. The total variance $\hat{V}$ is calculated by combining the within-chain variance with the between-chain variance (B/N), using the formula:

$$\hat{V} = \frac{N-1}{N}W + \frac{B}{N}, \tag{5.43}$$

where B is the variance of the chain means across multiple chains, and N is the length of each chain. If $\hat{R}$ is close to 1, it indicates that the chains have likely converged.

Effective Sample Size (NEFF), on the other hand, provides an estimate of the number of independent samples effectively obtained from the MCMC chain, accounting for the correlation between samples. The theoretical formula for NEFF is:

$$n_{\text{eff}} = \frac{MN}{1 + 2\sum_{t=1}^{\infty} \rho_t}, \tag{5.44}$$

where M is the number of chains, N is the number of samples per chain, and ρ_t is the autocorrelation at lag t. Since high autocorrelation reduces the amount of new information obtained from each sample, a larger sum of autocorrelations implies a smaller effective sample size.

I calculate both PSRF and NEFF for a collection of MCMC simulations following the methods from Brooks and Gelman (1998) and Gelman et al. (2013). I first split each

chain in half to mitigate the effects of non-stationarity. To compute the within-chain variance (W), I center each chain by subtracting its mean and summing the squared differences, normalising by $(N-1) \times M$. The between-chain variance (B) is estimated by calculating the variance of the means across the chains to quantify the variability due to differences between chains. The total variance estimate ($\hat{V}$) is then computed as a mixture of the within-chain and between-chain variances. Using these components, PSRF ($\hat{R}$) is computed as the square root of the ratio of $\hat{V}$ to W.

For NEFF estimation, I use variogram estimates to approximate the autocorrelation (ρ_t) at different lags. Specifically, Geyer's initial positive sequence is used to determine which lags of the autocorrelation sequence contribute positively. This approach yields a conservative estimate of the autocorrelation time (τ), which is used to calculate the effective sample size. The final NEFF is computed by dividing the total number of samples ($M \times N$) by the estimated autocorrelation time τ. Additionally, the method provides an estimated thinning factor based on Geyer's initial positive sequence, which can be used to determine how to thin the chains effectively if necessary.

The remainder of this section discusses the convergence for the two estimated models. For the one-type model, Table 5.12 describes these two statistics for each set of estimated parameters. $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\mu}$ all show clear convergence, with a highest PSRF of 1.002 and a lowest effective sample size of 4,557. The convergence of the correlation parameters, $\boldsymbol{\Sigma}$, is somewhat worse. This is due to high autocorrelation of the chains, as indicated by the low effective sample size. However, these results do not appear to indicate severe issues with the likelihood function and its underlying geometry that would require further exploration in Stan.

Table 5.12: Convergence of the one-type model

Parameters	Median PSRF	Mean PSRF	Highest PSRF
$\boldsymbol{\beta}$	1.000	1.000	1.001
$\boldsymbol{\gamma}$	1.000	1.000	1.001
$\boldsymbol{\Sigma}$	1.005	1.009	1.018
$\boldsymbol{\mu}$	1.000	1.000	1.002
Parameters	**Median NEFF**	**Mean NEFF**	**Lowest NEFF**
$\boldsymbol{\beta}$	7,904	7,725	6,913
$\boldsymbol{\gamma}$	7,565	7,500	6,623
$\boldsymbol{\Sigma}$	985	1,186	409
$\boldsymbol{\mu}$	7,990	7,913	4,557

Notes: This table reports convergence statistics for the estimated chains, including the potential scale reduction factor (PSRF) and number of effective samples (NEFF) for the beta parameters, cut-off parameters, Cholesky parameters ($\boldsymbol{\Sigma}$), and individual management quality parameters $\boldsymbol{\mu}$.

For the two-type model using country indicators as covariates, I run four parallel chains for 1,000 iterations. The convergence statistics are reported below in Table 5.13. For this model, prior to calculating the convergence diagnostics and reporting parameter estimates, I manually re-label the Markov chains based on the sign of the estimated α parameter. Since the mixture model is non-identifiable—meaning that the model remains equivalent if the labels for type 1 and type 2 are swapped—I impose an identification restriction by consistently assigning type 1 to represent the "MDB" type and type 2 to represent the "MAT" type. This re-labeling ensures consistent interpretation of the model parameters across chains and appropriate calculation of the convergence statistics.

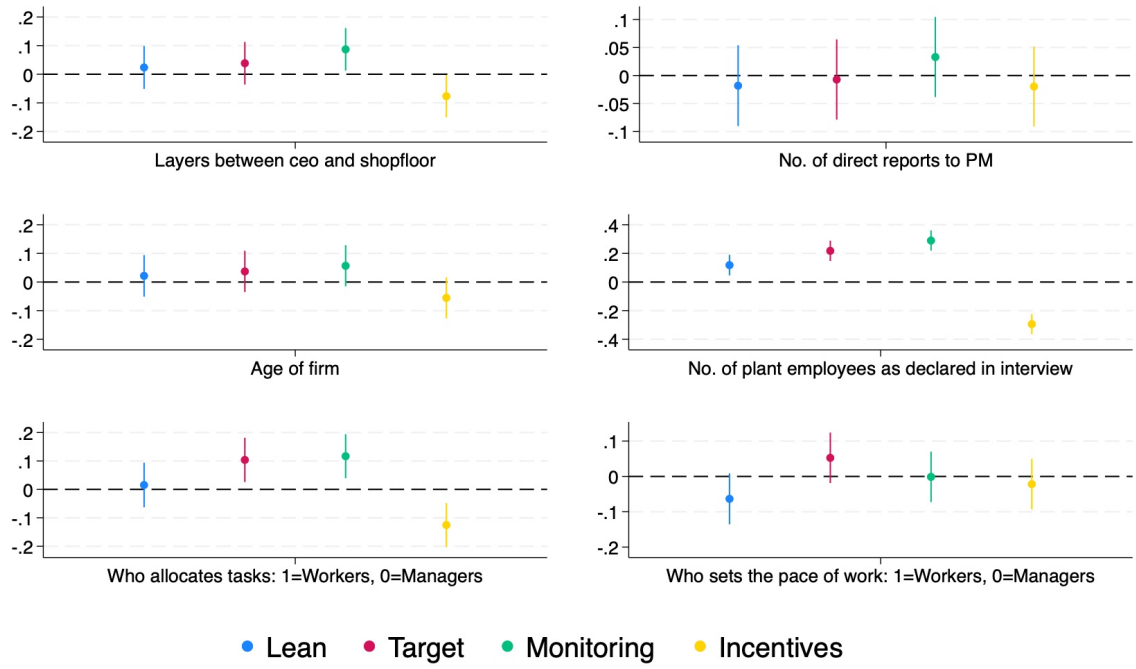
Table 5.13: Convergence of the two-type model

Parameters	**Min PSRF**	**Median PSRF**	**Mean PSRF**	**Max PSRF**
Beta	0.99	1.00	1.00	1.00
Gamma	1.00	1.00	1.00	1.00
Alpha	1.01	1.01	1.01	1.01
Delta	1.00	1.01	1.01	1.03
Mu	1.00	1.00	1.00	1.03
Cholesky	1.00	1.00	1.00	1.02
Parameters	**Min NEFF**	**Median NEFF**	**Mean NEFF**	**Max NEFF**
Beta	3,415	3,833	3,830	4,398
Gamma	3,023	3,587	3,555	3,890
Alpha	571	571	571	571
Delta	351	631	728	1,561
Mu	301	3,586	3,182	5,088
Cholesky	378	2,768	2,315	4,285

Notes: This table reports the convergence diagnostics for the two-type model with country fixed effects. The table presents the Potential Scale Reduction Factor (PSRF) and the Effective Sample Size (NEFF). The minimum, median, mean, and maximum PSRF values are reported for each parameter. The model has largely converged appropriately.

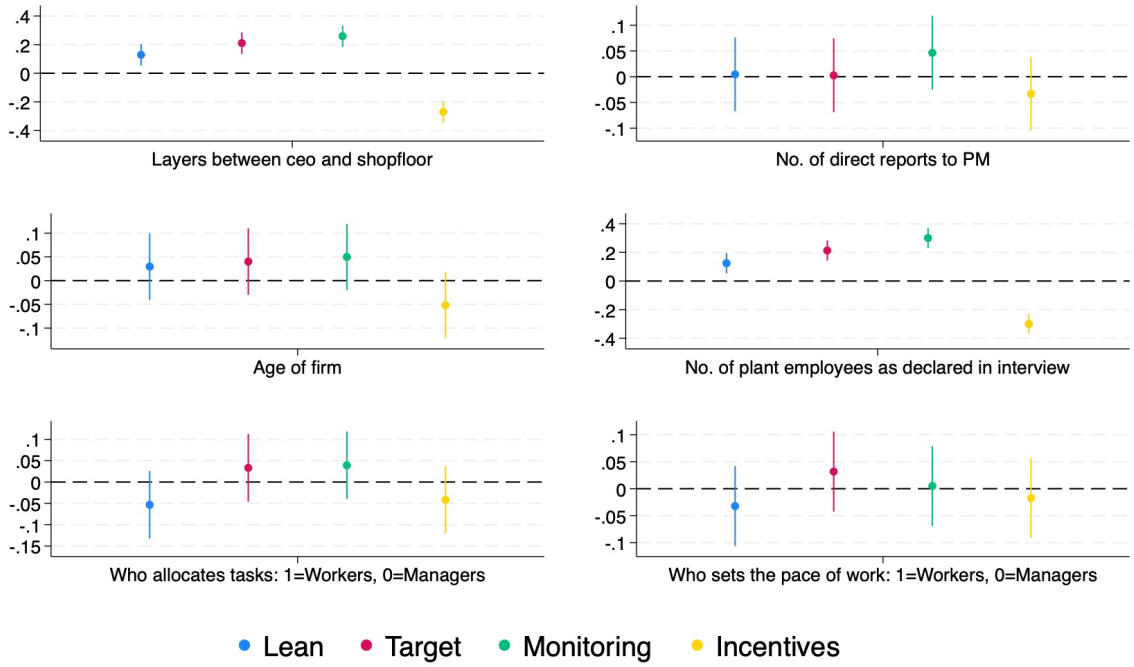
5.7.F Robustness

Figure 5.3: Latent management scores and firm observables with ownership fixed effects



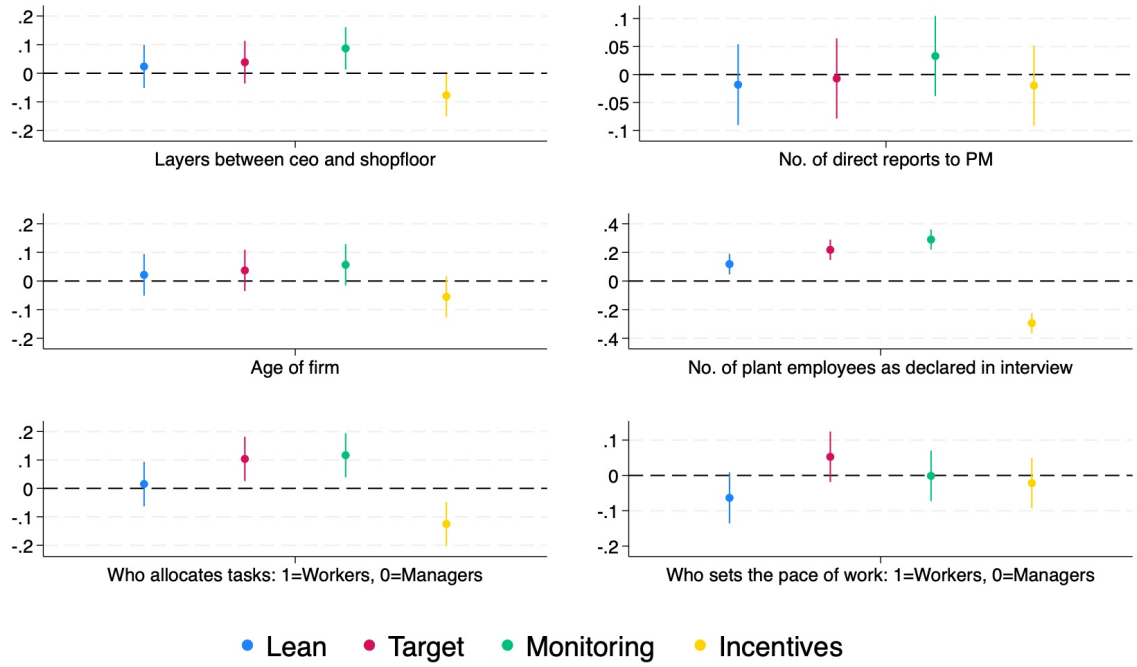
Notes This table of the score for each of the four dimensions on a dummy for a binary split of firms by layers between the ceo and shopfloor, number of direct reports to the PM, age of the firm, no. of employees, who allocates tasks and who sets the pace of work. This specification includes ownership fixed effects. All dependent variables are normalised to have mean zero variance one.

Figure 5.4: Latent management scores and firm observables with country fixed effects



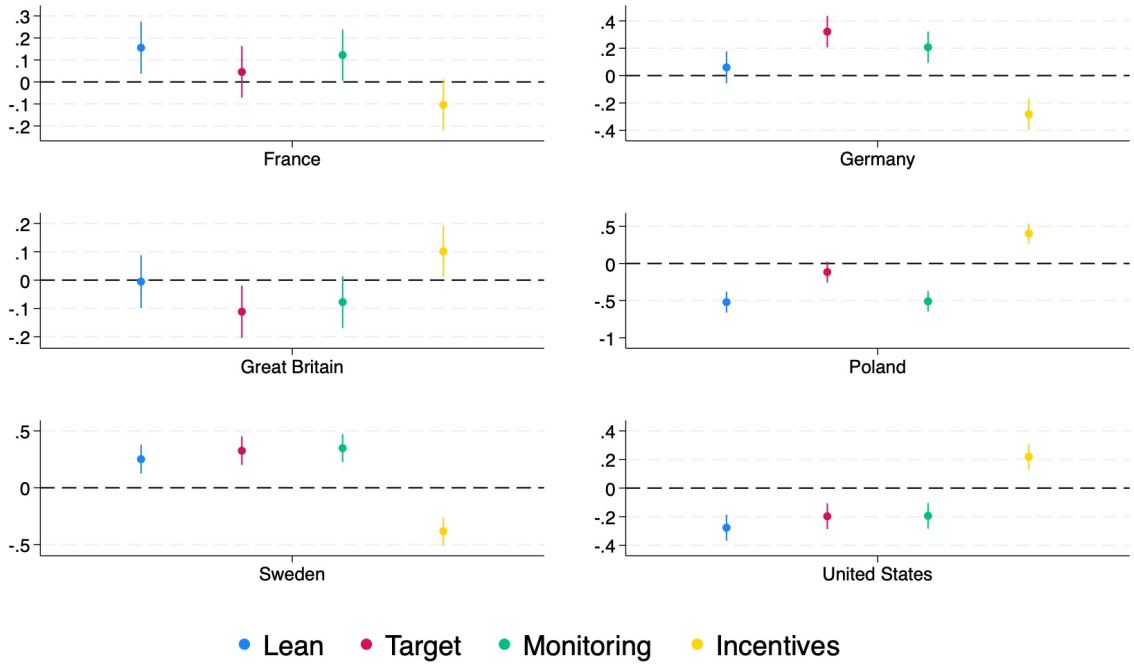
Notes This table of the score for each of the four dimensions on a dummy for a binary split of firms by layers between the ceo and shopfloor, number of direct reports to the PM, age of the firm, no. of employees, who allocates tasks and who sets the pace of work. This specification includes country fixed effects. All dependent variables are normalised to have mean zero variance one.

Figure 5.5: Latent management scores and firm observables with industry fixed effects



Notes This table of the score for each of the four dimensions on a dummy for a binary split of firms by layers between the ceo and shopfloor, number of direct reports to the PM, age of the firm, no. of employees, who allocates tasks and who sets the pace of work. This specification includes industry fixed effects. All dependent variables are normalised to have mean zero variance one.

Figure 5.6: Latent management scores and country dummies with industry and ownership fixed effects



Notes This table shows the OLS coefficient of a regression of the deviation from the expectation of the management score in each dimension of management on a country dummy for six countries. This specification includes industry and ownership fixed effects. All dependent variables are normalised to have mean zero variance one.

5.7.G Correlation between raw management scores by country

Table 5.14: Correlation between management scores by country

	France				Germany			
	Lean	Targets	Mon.	People	Lean	Targets	Mon.	People
Lean	1				1			
Targets	0.45	1			0.55	1		
Monitoring	0.53	0.66	1		0.62	0.65	1	
People	0.36	0.60	0.51	1	0.53	0.60	0.58	1
	Greece				Great Britain			
	Lean	Targets	Mon.	People	Lean	Targets	Mon.	People
Lean	1				1			
Targets	0.56	1			0.49	1		
Monitoring	0.68	0.72	1		0.60	0.68	1	
People	0.45	0.58	0.57	1	0.41	0.61	0.55	1
	Italy				Northern Ireland			
	Lean	Targets	Mon.	People	Lean	Targets	Mon.	People
Lean	1				1			
Targets	0.36	1			0.80	1		
Monitoring	0.48	0.65	1		0.70	0.52	1	
People	0.25	0.52	0.46	1	0.56	0.52	0.38	1
	Poland				Portugal			
	Operations	Targets	Mon.	People	Lean	Targets	Mon.	People
Lean	1				1			
Targets	0.40	1			0.40	1		
Monitoring	0.60	0.55	1		0.34	0.82	1	
People	0.36	0.53	0.53	1	0.28	0.55	0.46	1
	Sweden				United States			
	Lean	Targets	Mon.	People	Lean	Targets	Mon.	People
Lean	1				1			
Targets	0.38	1			0.46	1		
Monitoring	0.54	0.65	1		0.56	0.71	1	
People	0.43	0.53	0.55	1	0.34	0.51	0.49	1

Notes This table describes the correlation matrices between raw management scores for the 10 countries in alphabetical order.

Conclusion

This thesis explores several dimensions of organisational practices in medium-sized and large enterprises, focusing on low-income settings such as Ethiopia. The research provides new insights into the factors that influence firm performance, managerial decision-making processes, and the interaction between institutional structures and organisational outcomes.

In the first chapter, I develop a Bayesian adaptive questionnaire to elicit and quantify the use of formal and informal managerial rules. This tool enables a detailed analysis of HR managers' heuristic decision-making processes and the use of rules by managers. The findings reveal that firms with more rule-based management practices experience higher profitability but are less resilient during economic shocks, such as the COVID-19 pandemic. This suggests a trade-off between efficiency and adaptability, highlighting the dual nature of formal rules: while they foster consistency and profitability, they also introduce rigidities that reduce adaptability. Future work based on this could explore causal links between organisational rules and firm performance, examine how shared understanding and enforcement of rules affect firms, and investigate the impact of rules on resilience to economic shocks.

In the second chapter, the focus shifts to the supply and demand of managerial traits in Ethiopia, using an incentivised matching experiment. This experiment provides insights into how aspiring managers and firms value different managerial traits. By identifying distinct managerial types, we find that firms prefer managers who rely on formal authority and organisational policies. Additionally, the analysis suggests that labour market exposure has a causal influence on managerial traits, in particular helping individuals from a

more disadvantaged socioeconomic background adopt a style of management valued by potential employers – highlighting a novel potential channel of labour market exclusion. These findings suggest that targeted interventions could shape managerial traits to better align with firm needs. Future research could explore how labour market interventions affect managerial development, examine the performance of different managerial types within firms, and assess how diverse teams work together effectively.

In the third chapter, I examine how organisational values and social image concerns affect decision-making processes, particularly regarding equality of opportunity. Through a field experiment, I observe that reinforcing decision-makers' identities as organisational representatives improves the quality and consistency of their evaluations. This chapter underscores the importance of institutional design in aligning individual actions with broader organisational objectives. Future research could investigate how combining light-touch institutional interventions with more robust mechanisms, such as financial incentives or long-term diversity commitments, could mitigate systemic biases and foster greater inclusivity in decision-making.

In the final chapter, I explore the broader debate on management as a technology versus management as design. By applying a new empirical model to data from the World Management Survey, the chapter examines whether management practices should be seen as universally good or as contingent on organisational environments. The findings suggest that while there is often a positive correlation across different management practices, there are significant differences in how these practices interact depending on the country and organisational context. This points to the importance of a nuanced approach to understanding management, acknowledging that the effectiveness of practices may depend heavily on the specific setting in which they are implemented. The results and developed empirical methodology suggest several avenues for future research. First, using the World Management Survey, future work could examine the link between clusters of management practices, productivity, and organisational characteristics, as well as how these relationships change over time using the panel dimension of the dataset. Additionally, leveraging policy changes or IV approaches could provide causal evidence of what drives firms to

adopt specific combinations of practices. Given the range of sectors in which the World Management Survey is currently being implemented, including banking, healthcare, and education, there is significant scope to further explore these topics.

In summary, this thesis aims to contribute to a deeper understanding of the organisational dynamics that shape firm performance and managerial behaviour in low-income contexts. By developing innovative tools and applying them across a range of settings, this research provides the basis for future studies to better understand organisational management, practices, and policy interventions aimed at enhancing productivity and inclusivity in organisations.

References

Abebe, G., S. Caria, M. Fafchamps, P. Falco, S. Franklin, and S. Quinn (2021). Anonymity or distance? job search and labour market exclusion in a growing african city. *Review of Economic Studies 88*(3), 1279–1310.

Abebe, G., M. Fafchamps, M. Koelle, and S. Quinn (2019). Learning management through matching: A field experiment using mechanism design. Technical report, National Bureau of Economic Research.

Abebe, G., M. Fafchamps, M. Koelle, and S. Quinn (2024). Matching, management and employment outcomes: A field experiment with firm internships. *Working paper*.

Aghion, P., N. Bloom, B. Lucking, R. Sadun, and J. Van Reenen (2021). Turbulence, firm decentralization, and growth in bad times. *American Economic Journal: Applied Economics 13*(1), 133–169.

Ai, W., R. Chen, Y. Chen, Q. Mei, and W. Phillips (2016). Recommending teams promotes prosocial lending in online microfinance. *Proceedings of the National Academy of Sciences 113*(52), 14944–14948.

Akcigit, U., H. Alp, and M. Peters (2021a). Lack of selection and limits to delegation: firm dynamics in developing countries. *American Economic Review 111*(1), 231–275.

Akcigit, U., H. Alp, and M. Peters (2021b, January). Lack of selection and limits to delegation: Firm dynamics in developing countries. *American Economic Review 111*(1), 231–75.

Akerlof, G. A. and R. E. Kranton (2000). Economics and identity. *The quarterly journal of economics 115*(3), 715–753.

Alesina, A., S. Hohmann, S. Michalopoulos, and E. Papaioannou (2021). Intergenera-

tional mobility in africa. *Econometrica 89*(1), 1–35.

Alonso, R. and N. Matouschek (2008). Optimal delegation. *The Review of Economic Studies 75*(1), 259–293.

Armstrong, M. and J. Vickers (2010). A model of delegated project choice. *Econometrica 78*(1), 213–244.

Asch, S. E. (1951). Effects off Group Pressure Upon the Modification and Distrotion of Judgments. In H. S. Guetzkow (Ed.), *Groups, Leadership, and Men*, pp. 177–190. Carnegie Press.

Asher, S., P. Novosad, and C. Rafkin (2022). Intergenerational mobility in education in india. *American Economic Journal: Applied Economics 14*(2), 213–241.

Ashraf, N., O. Bandiera, and B. K. Jack (2014). No margin, no mission? a field experiment on incentives for public service delivery. *Journal of Public Economics 120*, 1–17.

Atkin, D., A. K. Khandelwal, and A. Osman (2019). Measuring productivity: Lessons from tailored surveys and productivity benchmarking. In *AEA Papers and Proceedings*, Volume 109, pp. 444–449. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codina (2015). Estimating the production function for human capital: Results from a randomized control trial in colombia. Working Paper 20965, National Bureau of Economic Research.

Autor, D. H., F. Levy, and R. J. Murnane (2002). Upstairs, downstairs: Computers and skills on two floors of a large bank. *ILR Review 55*(3), 432–447.

Bandiera, O., I. Barankay, and I. Rasul (2010). Social incentives in the workplace. *The Review of Economic Studies 77*(2), 417–458.

Bandiera, O., L. Guiso, A. Prat, and R. Sadun (2015). Matching firms, managers, and incentives. *Journal of Labor Economics 33*(3), 623–681.

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

Barley, S. R. (1986). Technology as an occasion for structuring: Evidence from observations of ct scanners and the social order of radiology departments. *Administrative Science Quarterly 31*(1), 78–108.

Barrios-Fernández, A., C. Neilson, and S. Zimmerman (2024, August). Elite universities and the intergenerational transmission of human and social capital. Working paper.

Bassi, V., J. H. Lee, A. Peter, T. Porzio, R. Sen, and E. Tugume (2023). Self-employment within the firm. Technical report, National Bureau of Economic Research.

Bassi, V. and A. Nansamba (2022). Screening and signalling non-cognitive skills: experimental evidence from uganda. *The Economic Journal 132*(642), 471–511.

Battaglia, L., T. Christensen, S. Hansen, and S. Sacher (2024). Inference for regression with variables generated by ai or machine learning.

Benmelech, E. and C. Frydman (2015). Military ceos. *Journal of Financial Economics 117*(1), 43–59.

Benson, A. M. and K. L. Shaw (2025). What do managers do? an economist's perspective. *NBER Working Paper 33431*.

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

Bianchi, N. and M. Giorcelli (2022). The dynamics and spillovers of management interventions: Evidence from the training within industry program. *Journal of Political Economy 130*(6), 1630–1675.

Biderman, D., D. Blei, W. Cai, A. Ciccarone, A. Feder, and A. Prat (in progress). Fixed effects topic model: Unsupervised identification of organizational culture from job text reviews. Working paper.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research 3*(Jan), 993–1022.

Bloom, N., E. Brynjolfsson, L. Foster, R. Jarmin, M. Patnaik, I. Saporta-Eksten, and J. Van Reenen (2019, May). What drives differences in management practices? *American Economic Review 109*(5), 1648–83.

Bloom, N., B. Eifert, A. Mahajan, D. McKenzie, and J. Roberts (2013a). Does manage-

ment matter? evidence from india. *The Quarterly journal of economics 128*(1), 1–51.

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

Bloom, N., C. Genakos, R. Sadun, and J. Van Reenen (2012a). Management practices across firms and countries. *Academy of management perspectives 26*(1), 12–33.

Bloom, N., C. Genakos, R. Sadun, and J. Van Reenen (2012b). Management practices across firms and countries. *Academy of Management Perspective 26*(1), 12–33.

Bloom, N., R. Lemos, R. Sadun, D. Scur, and J. V. Reenen (2014). Jeea-fbbva lecture 2013: The new empirical economics of management. *Journal of the European Economic Association 12*(4), 835–876.

Bloom, N., A. Mahajan, D. McKenzie, and J. Roberts (2010). Why do firms in developing countries have low productivity? *American Economic Review 100*(2), 619–623.

Bloom, N., R. Sadun, and J. Van Reenen (2016a). *Management as a Technology?*, Volume 22327. National Bureau of Economic Research Cambridge, MA.

Bloom, N., R. Sadun, and J. Van Reenen (2016b, May). Management as a technology? Working Paper 22327, 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*(4), 1351–1408.

Borgschulte, M., M. Guenzel, C. Liu, and U. Malmendier (2021, March). Ceo stress, aging, and death. Working Paper 28550, National Bureau of Economic Research.

Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education*, pp. 241–258. New York: Greenwood.

Bowles, S. and H. Gintis (1977). Schooling and the abandonment of the socialization hypothesis in economic theory. *The Review of Economics and Statistics 59*(4), 487–501.

Bowles, S. and H. Gintis (2002). Schooling in capitalist america revisited. *Scandinavian Journal of Economics 104*(1), 1–24.

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

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*(2), 635–687.

Brynjolfsson, E., S. Aral, and L. Wu (2010). Three-way complementarities: Performance pay, hr analytics and information technology. *Manage Sci 2010*, 1–33.

Brynjolfsson, E. and P. Milgrom (2013). Complementarity in organizations. *The handbook of organizational economics*, 11–55.

Brynjolfsson, E., D. Rock, and C. Syverson (2017, November). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. Working Paper 24001, National Bureau of Economic Research.

Bursztyn, L., G. Egorov, I. Haaland, A. Rao, and C. Roth (2023). Justifying dissent. *The Quarterly Journal of Economics 138*(3), 1403–1451.

Callander, S. and N. Matouschek (2019). The risk of failure: Trial and error learning and long-run performance. *American Economic Journal: Microeconomics 11*(1), 44–78.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2019, 11). Are Referees and Editors in Economics Gender Neutral?*. *The Quarterly Journal of Economics 135*(1), 269–327.

Caria, A. S., G. Gordon, M. Kasy, S. Quinn, S. O. Shami, and A. Teytelboym (2024). An adaptive targeted field experiment: Job search assistance for refugees in jordan. *Journal of the European Economic Association 22*(2), 781–836.

Caria, S., K. Orkin, and G. Bedoya (2023). Active labour market programmes. *VoxDevLit 7*(1).

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of statistical software 76*, 1–32.

Chang, X., L. Dai, L. Feng, J. Han, J. Shi, and B. Zhang (2025). A good sketch is better

than a long speech: evaluate delinquency risk through real-time video analysis. *Review of Finance 29*(2), 467–500.

Chapman, J., E. Snowberg, S. W. Wang, and C. Camerer (2024). Looming large or seeming small? attitudes towards losses in a representative sample. *Review of Economic Studies*, rdae093.

Charness, G. and P. Holder (2019). Charity in the laboratory: Matching, competition, and group identity. *Management Science 65*(3), 1398–1407.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2020). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report. Fischer-Shultz Lecture, Econometric Society World Congress, 2020.

Cohn, A., E. Fehr, B. Herrmann, and F. Schneider (2014). Social Comparison and Effort Provision: Evidence from a Field Experiment. *Journal of the European Economic Association 12*(4), 877–898.

Cunha, F., J. Heckman, and S. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. Working Paper 15664, National Bureau of Economic Research.

Dahlstrand, A., D. László, H. Schweiger, O. Bandiera, A. Prat, and R. Sadun (2025). Ceo-firm matches and productivity in 42 countries. Technical report, National Bureau of Economic Research.

DellaVigna, S., J. A. List, and U. Malmendier (2012, 01). Testing for Altruism and Social Pressure in Charitable Giving *. *The Quarterly Journal of Economics 127*(1), 1–56.

Dellavigna, S., J. A. List, U. Malmendier, and G. Rao (2016, 10). Voting to Tell Others. *The Review of Economic Studies 84*(1), 143–181.

Ellison, G. and R. Holden (2013, November). A theory of rule development. *The Journal of Law, Economics, and Organization 30*(4), 649–682.

Englmaier, F., J. Galdon-Sanchez, R. Gil, M. Kaiser, and H. Strandt (2020). Management practices and firm performance during the great recession: Evidence from spanish survey data. *Unpublished manuscript*.

Ennen, E. and A. Richter (2010). The whole is more than the sum of its parts—or is it? a review of the empirical literature on complementarities in organizations. *Journal of Management 36*(1), 207–233.

Ethiopian Statistical Service and World Bank (2023, September). Ethiopia socioeconomic panel survey (esps) report - wave 5, 2021/22. Published in September 2023.

Exley, C. L. and K. Nielsen (2024, March). The gender gap in confidence: Expected but not accounted for. *American Economic Review 114*(3), 851–85.

Fafchamps, M. and S. Quinn (2015, April). Aspire. Working Paper 21084, National Bureau of Economic Research.

Fafchamps, M. and S. Quinn (2018). Networks and manufacturing firms in africa: Results from a randomized field experiment. *The World Bank Economic Review 32*(3), 656–675.

Falk, A., F. Kosse, and P. Pinger (2020). Mentoring and schooling decisions: Causal evidence. (13387).

Feng, A., D. Lagakos, and J. Rauch (2024). Unemployment and development. *The Economic Journal 134*(658), 614–647.

Flammer, C. and J. Luo (2017). Corporate social responsibility as an employee governance tool: Evidence from a quasi-experiment. *Strategic Management Journal 38*(2), 163–183.

Frazier, P. (2018). A tutorial on bayesian optimization. *ArXiv abs/1807.02811*.

Garicano, L. and P. Heaton (2010). Information technology, organization, and productivity in the public sector: Evidence from police departments. *Journal of Labor Economics 28*(1), 167–201.

Garicano, L., I. Palacios-Huerta, and C. Prendergast (2005). Favoritism under social pressure. *Review of Economics and Statistics 87*(2), 208–216.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis 1*(3), 515 – 534.

Gelman, A. et al. (2013). Bayesian data analysis, 3rd: Boca raton. *Texts in Statistical Science*.

Gelman, A., D. Lee, and J. Guo (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics 40*(5), 530–543.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Gerber, A. S., D. P. Green, and C. W. Larimer (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *The American Political Science Review 102*(1), 33–48.

Gibbons, R. and R. Henderson (2012). Relational contracts and organizational capabilities. *Organization Science 23*(5), 1350–1364.

Giorcelli, M. (2019). The long-term effects of management and technology transfers. *American Economic Review 109*(1), 1–33.

Glick, P. and S. T. Fiske (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology 70*(3), 491–512.

Goleman, D. (2000). Leadership that gets results. *Harvard Business Review 78*(2), 78–89.

Gorodnichenko, Y., T. Pham, and O. Talavera (2023). The voice of monetary policy. *American Economic Review 113*(2), 548–584.

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

Griffiths, T. L. (2004). Finding scientific topics. *PNAS*.

Guenzel, M., S. Kogan, M. Niessner, and K. Shue (2025, January). Ai personality extraction from faces: Labor market implications. *SSRN Electronic Journal*.

Haaland, I., C. Roth, S. Stantcheva, and J. Wohlfart (2024). Measuring what is top of mind. *Working paper*.

Hall, B. H., J. Mairesse, and P. Mohnen (2010). Measuring the returns to r&d. In *Handbook of the Economics of Innovation*, Volume 2, pp. 1033–1082. Elsevier.

Handlan, A. and H. Sheng (2023). Gender and tone in recorded economics presentations:

Audio analysis with machine learning. *Working paper*.

Heckman, J. J. and Y. Rubinstein (2001). The importance of noncognitive skills: Lessons from the ged testing program. *American Economic Review 91*(2), 145–149.

Hoffman, M., L. B. Kahn, and D. Li (2018). Discretion in hiring. *The Quarterly Journal of Economics 133*(2), 765–800.

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

Hoffman, M. D., A. Gelman, et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res. 15*(1), 1593–1623.

Hsieh, C.-T. and B. A. Olken (2014, September). The missing "missing middle". *Journal of Economic Perspectives 28*(3), 89–108.

Ichniowski, C., K. Shaw, and G. Prennushi (1997). The effects of human resource management practices on productivity: A study of steel finishing lines. *American Economic Review 87*(3), 291–313.

Jedynak, B., P. I. Frazier, and R. Sznitman (2012). Twenty questions with noise: Bayes optimal policies for entropy loss. *Journal of Applied Probability 49*(1), 114–136.

Juhász, R., M. P. Squicciarini, and N. Voigtländer (2024). Technology adoption and productivity growth: Evidence from industrialization in france. *Journal of Political Economy 132*(10), 000–000.

Kaplan, S. N., M. M. Klebanov, and M. Sorensen (2012). Which CEO characteristics and abilities matter? *The Journal of Finance 67*(3), 973–1007.

Kasy, M. and A. Sautmann (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica 89*(1), 113–132.

Khan, M. Y. (2020). Mission motivation and public sector performance: experimental evidence from pakistan.

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

Lemos, R. and D. Scur (2018). All in the family? ceo choice and firm organization.

Li, D., L. R. Raymond, and P. Bergman (2020). Hiring as exploration. Working Paper 27736, National Bureau of Economic Research.

Li, J., A. Mukherjee, and L. Vasconcelos (2022, September). What makes agility fragile? a dynamic theory of organizational rigidity. Article 69(6), Management Science.

Li, S. (2017). Obviously strategy-proof mechanisms. *The American Economic Review 107*(11), 3257–3287.

López-Peña, P., M. Mozumder, A. Rabbani, and C. Woodruff (2025). Toxic Managers, Firm Productivity, and Worker Well-Being: Evidence from Bangladeshi Garment Factories. *CEPR Discussion Paper No. 19936.*.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: John Wiley & Sons.

Ludwig, J. and S. Mullainathan (2023). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics 139*(2), 751–827.

Macchiavello, R., A. Menzel, A. Rabbani, and C. Woodruff (2020, July). Challenges of change: An experiment promoting women to managerial roles in the bangladeshi garment sector. Working Paper 27606, National Bureau of Economic Research.

Malmendier, U., G. Tate, and J. Yan (2011). Overconfidence and early-life experiences: The effect of managerial traits on corporate financial policies. *The Journal of Finance 66*(5), 1687–1733.

McKenzie, D. and D. Sansone (2019). Predicting entrepreneurial success is hard: Evidence from a business plan competition in nigeria. *Journal of Development Economics 141*, 102369.

Meager, R. (2019, January). Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics 11*(1), 57–91.

Metcalfe, R. D., A. B. Sollaci, and C. Syverson (2023). Managers and productivity in retail. Technical report, National Bureau of Economic Research.

Milgrom, P. and J. Roberts (1994). Complementarities and systems: Understanding japanese economic organization. *Estudios Económicos 9*(1), 3–42.

Milgrom, P. and J. Roberts (1995). Complementarities and fit strategy, structure, and organizational change in manufacturing. *Journal of Accounting and Economics 19*(2), 179 – 208. Organizations, Incentives, and Innovation.

Neal, R. M. and R. M. Neal (1996). Monte carlo implementation. *Bayesian learning for neural networks*, 55–98.

Ouedraogo, R. and N. Syrichas (2021). *Intergenerational social mobility in Africa since 1920*. International Monetary Fund.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 24*(2), 193–202.

Porter, M. E. et al. (1996). What is strategy?

Press, I. P. (1989). Taiichi ohno, toyota production system (beyond. *Human Systems Management 8*, 175–182.

Rasmussen, C. E. and C. K. I. Williams (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Rasul, I. and D. Rogger (2018). Management of bureaucrats and public service delivery: Evidence from the nigerian civil service. *The Economic Journal 128*(608), 413–446.

Rivkin, J. W. (2000). Imitation of complex strategies. *Management science 46*(6), 824–844.

Scur, D., R. Sadun, J. Van Reenen, R. Lemos, and N. Bloom (2021). The world management survey at 18: lessons and the way forward. *Oxford Review of Economic Policy 37*(2), 231–258.

Shukla, S. (2025). Making the elite: Class discrimination at top firms. Working paper.

Siggelkow, N. (2001). Change in the presence of fit: The rise, the fall, and the renascence of liz claiborne. *Academy of Management Journal 44*.

Siggelkow, N. (2002a). Evolution toward fit. *Administrative Science Quarterly 47*(1), 125–159.

Siggelkow, N. (2002b). Misperceiving interactions among complements and substitutes: Organizational consequences. *Management Science 48*, 900–916.

Stan Development Team (2024). *Stan Reference Manual*. v2.36.0 https://mc-stan.org.

Stantcheva, S. (2021). Understanding tax policy: How do people reason? *The Quarterly Journal of Economics 136*(4), 2309–2369.

Steyvers, M., P. Smyth, M. Rosen-Zvi, and T. Griffiths (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 306–315.

Syverson, C. (2011). What determines productivity? *Journal of Economic literature 49*(2), 326–365.

Taylor, F. W. (1911). Principles and methods of scientific management. *Journal of Accountancy 12*(3), 3.

Teece, D. J., G. Pisano, and A. Shuen (1997). Dynamic capabilities and strategic management. *Strategic management journal 18*(7), 509–533.

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

Weidmann, B., J. Vecci, F. Said, D. J. Deming, and S. R. Bhalotra (2024, July). How do you find a good manager? Working Paper 32699, National Bureau of Economic Research.

Welz, M., A. Alfons, M. Demirer, and V. Chernozhukov (2022). *GenericML: Generic Machine Learning Inference*. R package version 0.2.2.

World Bank (2022). Ethiopia gender diagnostic: Building the evidence base to address gender inequality in ethiopia. Technical report, World Bank, Africa Gender Innovation Lab.

Zimmerman, S. D. (2019). Elite colleges and upward mobility to top jobs and top incomes. *American Economic Review 109*(1), 1–47.