Appendix 1

1 Proof of increasing costs of performance pay

Here we provide a proof for the claim that expected costs of performance pay are increasing in the degree of decentralization. More formally, we show that $\bar{w}'(z_A) > 0$.

$$\bar{w}(z_A) = E[w_A(z_A)|e = 1] = \\ (1 - k + q) \quad \cdot \exp\left\{U_R + \tfrac{k}{q} \cdot C \cdot z_A\right\} + (k - q) \cdot \exp\left\{U_R - \tfrac{1-k}{q} \cdot C \cdot z_A\right\}$$

Taking the derivative with respect to z_A gives

$$\bar{w}'(z_A) = \frac{C}{q}\left[(1 - k + q) \cdot k \cdot w_A^H - (1 - k) \cdot (k - q) \cdot w_A^L\right]$$

with $\bar{w}'(z_A) > 0$ if the term in square brackets is positive. This will be the case if the following condition is met

$$\frac{w_A^H}{w_A^L} > \frac{1 - k}{k}\left(\frac{k - q}{1 - k + q}\right)$$

Note that as $q \in (0, 1)$ and $k \in (0, 1)$, the expression on the right hand side is always less than one. Also, $w_A^H > w_A^L$, so this inequality will always hold, ensuring that expected costs of performance pay increase in z_A.

2 Model Extension: Three decision layers in the hierarchy

2.1 Overview

In this section we present an extension of our baseline model of task allocation and performance pay to three layers. This extended setup helps to conceptually clarify our key empirical finding of concentration of decision authority at the managerial level, instead of a general movement towards more decentralization throughout the entire organization.

2.2 Technology and Preferences

There are now three levels in the hierarchy: the principal (P), the managerial employee (M) and the non-managerial employee (N). As before, the index $z \in [0, 1]$ denotes problem complexity, with higher values

of z denoting a more complex problem.

Decentralization

Task allocation across decision layers is captured by two decentralization cutoff points. $z_M > 0$ is the problem complexity cutoff point that applies to managers. $z_N \in [0, z_M]$ is the problem cutoff point that applies to non-managerial employees.

Production and effort

As in our baseline setup, managers and non-managerial employees exert hidden effort $e \in \{0, 1\}$, which influences the probability of successfully solving a problem. Let $x_l \in \{L, H\}$ denote functions that summarize whether a problem was solved by agent $l \in \{N, M\}$. In the high output state $x_l = H$, the problem has been solved and output is one. In the low output state $x_l = 0$, the problem could not be solved and output is zero. Success probabilities or fractions of solved problems are denoted by

$$p_M^0 = P\left(x_M = H \middle| e_M = 0\right), \quad p_M^1 = P\left(x_M = H \middle| e_M = 1\right), \quad q_M = p_M^1 - p_M^0 \tag{1}$$

where p_M^1 is the probability managers solve a given problem if they exert effort and q_M is the incremental increase in the probability of success if managers exert effort. Similarily for non-managers, we can define

$$p_N^0 = P\left(x_N = H \middle| e_N = 0\right), \quad p_N^1 = P\left(x_N = H \middle| e_N = 1\right), \quad q_N = p_N^1 - p_N^0 \tag{2}$$

Combining decentralization decisions and success probabilities has the following implications for production:

- Principals are solving all problems with probability one and are assigned all problems that managers cannot solve. Therefore, production from principals is given by $1 - F(z_M)$
- Managers are assigned tasks that are not solved by principals (below z_M) and are not decentralized to non-managers (above z_N). Combining these two task allocation decisions, the mass of problems assigned to managers is $F(z_M) - F(z_N)$, multiplied by the probability that managers can solve the assigned problem. In other words, production is given by: $[F(z_M) - F(z_N)] \cdot P\left(x_M = H \middle| e_M\right)$
- Non-managers are assigned the simplest problems that are not assigned to managers and solve them with a probability of success that is a function of their effort. This implies that the non-managerial contribution to production is given by $F(z_N) \cdot P\left(x_N = H \middle| e_N\right)$

Effort costs

Similar to our original model, effort costs for agent $l \in \{N, M\}$ can be divided into two types of costs.

- Observable effort costs: $a_{1,l} z_l$, which can be thought of training costs or equipment costs, which rise in the problem complexity assigned to agent $l \in \{N, M\}$
- Unobservable effort costs: $a_{2,l} z_l$, which capture the idea that more complex problems require agents to exert more effort. These effort costs are assumed to increase in the complexity of the most complicated problem agents are confronted with.

2.3 Optimal performance pay contract

The optimal contracting problem, given decentralization decisions z_N, z_M for agent $l \in \{N, M\}$ is given by

$$\min_{w_l^H, w_l^L} P\left(x_l = H \middle| e_l = 1\right) \cdot w_l^H + P\left(x_l = L \middle| e_l = 1\right) \cdot w_l^L \tag{3}$$

subject to:

$$\begin{aligned} \text{(IC)} \quad & E\left[U_l\left(w_l, z_l, e_l = 1\right) \middle| e_l = 1\right] \geq E\left[U_l\left(w_l, z_l, e_l = 0\right) \middle| e_l = 0\right] \\ \text{(IR)} \quad & E\left[U_l\left(w_l, z_l, e_l = 1\right) \middle| e_l = 1\right] \geq U_l^R \,. \end{aligned}$$

Similar to our original model, we assume that preferences are given by

$$U_l(w_l, e_l) = \ln(w_l) - e_l \cdot a_{2,l} \cdot z_l \tag{4}$$

Combining preferences (4) with success probabilities (1) or (2) in the optimal contracting problem (3), gives the optimal performance pay contract for agent $l \in \{N, M\}$ as a function of decentralization:

$$\bar{w}_l(z_l) = E[w_l(z_l)|e_l = 1] = p_l^1 w_l^H + (1 - p_l^1) w_l^L \tag{5}$$

with optimal compensation as a function of output states given by

$$\begin{aligned} w_l^H &= \exp\left\{U_l^R + \frac{1 - p_l^0}{q_l} a_{2,l} z_l\right\} \\ w_l^L &= \exp\left\{U_l^R - \frac{p_l^0}{q_l} a_{2,l} z_l\right\} \end{aligned}$$

Two properties are noteworthy for the structure of optimal performance pay contracts.

First, as shown in Section 1 of this appendix, it is true that $\bar{w}_l'(z_l) > 0$. In words, the costs of performance pay increase in the degree of decentralization.

Second, how strong the performance-related component of compensation is can be measured by the difference between compensation when an employee solves a problem and when a problem is not solved, or $w_l^H - w_l^L$. Since managers will typically solve harder problems than non-managers, $z_M > z_N$, this directly implies that all other things equal, the performance-related component of compensation is stronger for managers than non-managers. If we could measure overall compensation levels, we would therefore expect that performance pay plays a more important role in overall compensation for managers than for non-managers.

We do not test this empirical implication in our study, as we do not have data on the contribution of performance pay to overall compensation for managers and non-managers. However, we note that the model shows that the fact that performance pay plays a bigger role in compensation for managers than non-managers is not an alternative explanation for our main empirical result that decision authority is concentrated at the manager level for firms that adopt performance pay. The model makes clear that even if one were to observe that performance pay incentives are stronger for managers than non-managers, the underlying mechanism is driven by differences in the complexity of problems. Managers are solving more complicated problems, and hence their unobservable effort cost is higher than the unobservable effort cost for non-managers. Stronger performance pay incentives for managers does not naturally lead to concentration of tasks at the managerial level of the hierarchy.

2.4 Optimal Decentralization without Performance Pay

$$\max_{\{z_M, z_N\}} \Pi\left(z_M, z_N, 0\right) = \tag{6}$$

$$\begin{aligned} [1 - F(z_M)]\cdot \quad & (1 - h \cdot a_P) \\ + \quad & p_M^0 \left[\cdot F(z_M) - F(z_N)\right] - (a_{1,M} + a_{1,N}) z_M \\ + \quad & \left(p_N^0 - p_M^0\right) \cdot F(z_N) - a_{1,N} z_N \end{aligned}$$

The first line captures the production contribution from the principal P as in our baseline model. The second line describes the production from managers. The third line captures the production contribution from non-managers. Gathering terms, this problem becomes

$$\begin{aligned} \Pi\left(z_M, z_N, 0\right) \quad & = \\ [1 - F(z_M)]\cdot \quad & (1 - h \cdot a_P) \\ + \quad & p_M^0 \cdot F(z_M) - a_{1,M} z_M \\ + \quad & [p_N^0 - p_M^0] \cdot F(z_N) - (a_{1,N} - a_{1,M}) z_N \end{aligned}$$

which makes clarifies how the addition of the non-manager layer influences the nature of the decentralization problem. In particular, there is an additional decentralization margin z_N, which captures the problem allocation between managerial and non-managerial employees. Recall that higher levels of z_N imply more decentralization to non-managerial employees and allocation of problems away from managers. The third term shows the costs and benefits associated with further decentralization from managers to non-managers:

- **Benefits of decentralizing from managers to non-managers**
  To capture the benefits of a further decentralization of decision making from managers to non-managers, we assume that $p_N^0 - p_M^0 > 0$. In other words, if employees do not exert unobservable effort, non-managers are more likely to be able to solve a given problem than managers. This assumption reflects several advantages that non-managerial employees at the frontline of the organizations might have, including local information and familiarity with operation or implementation issues.

- **Costs of decentralizing from managers to non-managers**
  Balancing the advantages of decentralization are increased observable effort costs, formalized as $a_{1,N} - a_{1,M} > 0$. Our preferred interpretation is that these costs including training costs of non-managers that principals need to expend in case of decentralizing more decisions to non-managers.

Taking first order conditions of (6) with respect to z_N, z_M yields the optimal decentralization decisions

without performance pay. These are given by

$$f\left(z_M^{*,0}\right) = \frac{a_{1,M}}{ha_P - (1 - p_M^0)} \tag{7}$$

$$f\left(z_N^{*,0}\right) = \frac{a_{1,N} - a_{1,M}}{p_N^0 - p_M^0} \tag{8}$$

2.5 Optimal Decentralization with Performance Pay

$$\max_{\{z_M, z_N\}} \Pi\left(z_M, z_N, 1\right) = \tag{9}$$

$$\begin{aligned}
[1 - F(z_M)]\cdot \quad & (1 - h \cdot a_P) \\
+ \quad & p_M^1 \left[\cdot F(z_M) - F(z_N)\right] - (a_{1,M} + a_{1,N}) z_M - \bar{w}_M(z_M) \\
+ \quad & \left(p_N^1 - p_M^1\right) \cdot F(z_N) - a_{1,N} z_N - \bar{w}_N(z_N)
\end{aligned}$$

As in our baseline model, the presence of performance pay influences the decentralization decision of the principal through two distinct channels. First, the performance pay contracts for employees ensure that hidden effort are exerted, so that probability of successfully solving problems are given by p_N^1, p_M^1, rather than p_N^0, p_M^0. Second, since decentralization influences the costs of hidden effort, it also influences the expected performance pay costs $\bar{w}_N(z_N), \bar{w}_M(z_M)$. The extended three layer model takes account of the fact that further decentralization of decisions from managers to non-managers will also influence performance pay of non-managers.

The optimal problem allocation decisions for managers and non-managers are given by:

$$f\left(z_M^{*,1}\right) = \frac{a_{1,M} + \bar{w}'_M(z_M)}{ha_P - (1 - p_M^0)} \tag{10}$$

$$f\left(z_N^{*,1}\right) = \frac{(a_{1,N} - a_{1,M}) + \bar{w}'_N(z_N)}{p_N^0 - p_M^0} \tag{11}$$

2.6 Implications of Performance Pay for Decentralization

There are two decentralization margins in the three layer model. The degree of decentralization from principal to managers is basically unchanged from our baseline model. Conditions under which the adoption of performance pay implies more decentralization to managers are obtained from combining (7) with (10) and

are given by

$$z_M^{*,1} > z_M^{*,0} \text{ if and only if } \frac{q_M}{h \cdot a_P - (1 - p_M^0)} > \frac{\bar{w}'_M\left(z_M^{*,1}\right)}{a_{1,M}} \tag{12}$$

To obtain conditions which imply that organizations with performance pay optimally centralize decisions from non-managers to managers, one needs to combin (8) with (11). These conditions are given by

$$z_N^{*,1} < z_N^{*,0} \text{ if and only if } \bar{w}'(z_N) > \left(\frac{p_N^1 - p_M^1}{p_N^0 - p_M^0} - 1\right) \cdot (a_{1,N} - a_{1,M}) \tag{13}$$

Notice that $\bar{w}'(z_N) > 0$ as well as $a_{1,N} - a_{1,M} > 0$. Put differently, if $\left(\frac{p_N^1 - p_M^1}{p_N^0 - p_M^0} - 1\right) < 0$, this condition will always be met. Therefore, the the key sufficient condition that implies that there is an optimal centralization of decisions from non-managerial to managerial employees is therefore given by

$$p_M^1 - p_M^0 > p_N^1 - p_N^0 \tag{14}$$

This is a supermodularity (or single-crossing) condition that captures the key complementarity that is needed to observe the concentration of decision making at the manager level. It states that the productivity effect of unobsered effort is increasing in the level of employee hierarchy. In other words, the incremental impact of effort on the probability to solve a problem is higher for managers than for non-managers.