Kirchmaier, Isadora

Conference Paper

Service Organizations: Customer Contact and Incentives of Knowledge Managers

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: Industrial Organization III, No. C11-V2

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Kirchmaier, Isadora (2014) : Service Organizations: Customer Contact and Incentives of Knowledge Managers, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: Industrial Organization III, No. C11-V2, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg

This Version is available at:
https://hdl.handle.net/10419/100418

Service Organizations: Customer Contact and Incentives of Knowledge Managers

Isadora Kirchmaier *

We analyze the interdependence of human resource management and knowledge management. The service organization is modeled as a queueing network. The optimal number of workers in each division, the amount of customer contact and the wage for each manager is determined. We combine three features within the model. First, each manager may engage in customer contact. We show that although the fraction of time a worker is busy is increasing in rank of the manager, the customer task acceptance rate is not necessarily monotonic. Second, knowledge management is explicitly taken into account. Knowledge acquired by workers depends on the effort of the manager. Third, since this effort is not easily measurable, a moral hazard problem might occur. We discuss a bonus contract under different performance evaluation schemes. If queueing costs increase we find it might be optimal to increase the knowledge and to decrease the number of workers. This implies that decisions are more decentralized. In a numerical example we analyze the elimination of middle management. A flattened firm may respond more quickly by pushing decisions downwards. However, we find that the mean response time is higher and the senior manager is more involved in internal tasks.

Subject classifications: organizational design; multi-agent moral hazard; queueing network

---

1 Introduction

Organizations provide a framework for employees to perform their tasks. The structure determines basic patterns such as who deals with incoming task from clients, who handles a task, which was already dealt with by another employee but could not be solved, and who decides that

*Address: Bergheimer Str. 58, 69115 Heidelberg, telephone: +49 6221 54 2951 , e-mail: i.m.kirchmaier@uni-heidelberg.de.

a task cannot be solved at all. It influences the incentives of each employee and thereby the flow of production and communication within an organization. The aim of human resource management (HRM) is to provide an environment so that the objectives of the organizations are accomplished through employees. The link between HRM and business outcome is well documented (Koys 2001, Paul and Anantharaman 2003, Purcell and Kinnie 2007, Bowen and Ostroff 2004). The specific goal of HRM is achieving high performance through people (Armstrong 2003). In order to achieve high performance through people, employees abilities, motivations and opportunities have to align, i.e. they need to have the necessary skills to perform the task, the employment contract has to provide adequate incentives and the work environment has to provide possibilities for expression also if a problem occurs (Boxall and Purcell 2003). Due to a shift from manufacturing-oriented to service organizations, customers are a source of production inputs. This introduces uncertainty in the task characteristics and thereby in the solvability of the task (Larsson and Bowen 1989). The employee in the division who deals with an incoming task may not have the knowledge to solve it. A popular way to address this lack of knowledge is knowledge management. The manager takes on the job of providing a knowledge-creating and sharing environment for the workers (Gao et al. 2008). Nonaka and Takeuchi (1995) define these managers responsible for the knowledge management as the knowledge-creating crew. They coach the employees thereby producing long-lasting learning, which guides the employees to enhance their performance (Redshaw 2000). However, managers differ in their willingness to coach. It is a time consuming task and might interfere with achieving performance targets (Goleman 2000). HRM provides the wage contract for the managers and thereby influences the incentive of the managers to coach their subordinates. Since the effort in coaching is not easily measurable, it might not be possible to contingent the contract on that effort and a moral hazard problem occurs.

The aim of this paper is to analyze the interdependence of the goals of HRM and the knowledge crew, i.e. the knowledge managers, who act as coaches for the employees in their division, taking the connectedness of the divisions through the supervisory structure into account. Especially, the reward structure, knowledge management, customer task acceptance rate and span of control of the knowledge managers are discussed. Furthermore, different information structures between HRM and the knowledge managers are considered. Under symmetric information, it is assumed that the amount of coaching is contractable while under asymmetric information HRM cannot observe the

effort of the knowledge manager. We ask the question if it is still possible for HRM to provide a bonus contract for the knowledge managers such that it is optimal to implement the amount of coaching under symmetric information.

Consider a service organization, which consists of different divisions and suppose that the organization generates revenue from solving tasks. The organization is modeled as a queueing network in which the divisions are the nodes and the workers are the servers of each division. The job of the workers is to solve tasks given by their knowledge manager. If the worker is busy, incoming tasks are queued. Tasks arrive due to customer contact of the knowledge manager and also, depending on the underlying supervisory structure, from knowledge managers of other divisions. For each division a knowledge manager is in charge of coordinating the incoming and outgoing tasks. If after processing, a division is not able to solve a task, the knowledge manager may forward it to other divisions of the organization. Also, the knowledge manager coaches the workers in his division, e.g. the workers in his span of control, and thereby influences the acquired knowledge of the worker. The organization's HRM is in charge of designing a wage contract for each knowledge manager. These contracts provide the incentive structure within the organization and influence the effort in coaching of the knowledge managers. Additionally, HRM pursues its other goals, determining the optimal amount of customer contact of the knowledge manager, i.e. the task acceptance rate, and the optimal number of workers in each division.

One example for such a type of organization is a consulting firm. A consulting firm typically consists of junior and senior managers where each manager supervises a division. Junior and senior managers receive tasks from customers and allocate them to their workers in the division. Tasks are queued, if all workers are busy. If the division of the junior manager is not able to solve a given tasks, the junior manager may ask the senior manager for help. An important part of the job description of each manger is coaching of the workers in his division. In an numerical example we consider such a type of organization and discuss the effect of an implementation of middle management.

There are several studies which view organizations as a network of queues. The one closest to our setup is Beggs (2001). He considers an organization, where depending on their rank, workers may differ in their ability to solve tasks. A task is queued until a worker is free to deal with it. If a worker is not able to solve a task he can send it to a worker on the next division. The wage is

a function of the contractable ability of each worker. The objective is to minimize wage costs but also to minimize network performance measures such as the average delay of a task. The trade-off between wage costs and delay determines the optimal ability level and number of workers in each division. Under the assumption that a task can be immediately dealt with Garicano (2000) derives the optimal organizational structure under symmetric information. He shows that there is one division of workers, that specializes in production and the other divisions specialize in supporting that division by attending to unsolved problems. The knowledge of the worker is increasing in the rank of the division. Calvo and Wellisz (1978) and Qian (1994) assumes that the lowest division is responsible for production. All the other divisions are supervisors, who invest resources in monitoring immediate subordinates. The output of each division will be used as an input for the next division. They show that if the divisions have the possibility to shirk without their supervisor knowing, then it is optimal to pay a wage increasing in rank, even if all workers have the same abilities.

We provide three main extensions. First, the production process is extended such that each division may accept tasks from customers in addition to unsolved tasks forwarded from subordinates. It provides additional insight, since we show that the optimal task acceptance rate is not necessarily monotonic in rank of the knowledge managers. The task acceptance rate of a knowledge manager depends on the task arrival rate and on the number of unsolved tasks forwarded by direct subordinate knowledge managers. In general, a knowledge manager of higher rank has a lower task acceptance rate. However, if a knowledge manager of lower rank receives comparably more unsolved problems it can occur that his task acceptance rate is lower than for the knowledge manager of higher rank. Nevertheless, the traffic intensity, which is the mean number of tasks in service at a worker, is increasing in the rank of the knowledge manager. In the numerical example we discuss the case, in which it is optimal that a middle manager has a lower task acceptance rate than the senior manager. The middle manager is mainly in charge of dealing with unsolved tasks of junior managers.

Second, we take knowledge management explicitly into account. The amount of knowledge acquired by workers depends on the effort of the knowledge crew. The knowledge managers are responsible for the creation and circulation of knowledge in their division. Knowledge creation and coaching induces costs for the knowledge manager. We show that the optimal variable costs of

knowledge creation per worker incurred by the knowledge manager of each division is increasing in the rank of the manager.

Third, since the effort in coaching is not easily measurable, we ask the question if it is still possible for HRM to provide a bonus contract for the knowledge managers such that it is optimal to implement the amount of coaching under symmetric information. Since knowledge creation and coaching induces costs for the knowledge manager, HRM has to take these aspects into account when designing a contract for a knowledge manager. Organizational design incorporates also control tools such as reward structures, task characteristics, and information systems. In a field study of compensation practices, Eisenhardt (1985) shows that task characteristics and the available information system to measure outcomes is strongly related to the choice of the reward structure. If the characteristic of a given task is not known in advance and the outcome is measurable, then it is more likely that the reward structure is based on the outcome and not on the behaviour of the manager. We analyze two different types of reward structures for the knowledge managers. Under independent performance evaluation only the own output influences the payment of the bonus, while under joint performance evaluation the performance of other divisions are relevant as well. We identify conditions, which depend crucially on the hazard rate of effort, under which a bonus contract exists. The hazard rate of effort gives the change in the cumulative distribution of output resulting from higher effort in relation to the probability to receive a bonus. We find that under independent performance evaluation the bonus is unambiguously positive, while under joint performance evaluation instead of a bonus payment a penalty could occur. A reason is that under joint performance evaluation an increase in knowledge of a worker has two effects. First, it increases the expected output of that division. Second, it decreases the output of the divisions, which supervise that division, since less unsolved tasks are forwarded. If the second effect outweighs the first effect, it is less likely that a certain overall output is achieved. Consider the senior knowledge manager, i.e. the manager who is not supervised by any other knowledge managers. If a contract under joint performance evaluation exists, it is optimal that the senior manager receives a bonus payment and not a penalty, since the second effect does not occur. We find that for the same target output, the wage sensitivy to performance for the senior manager is higher under joint performance evaluation than under individual performance evaluation. The wage sensitivity to performance is the ratio of the bonus payment to the salary plus bonus payment and is a common measure in

accounting (Baiman et al. 1995).

Within this framework we discuss the reaction of HRM to a change in the intensity of queueing costs. Beggs (2001) shows that if a delay becomes more costly, the ability of the employees increases. We also find that the knowledge of the workers and the number of workers in each division are substitutes. If delay becomes more costly, there are three effects on the number of workers in each division. First, since workers are able to solve more problems, less workers will be employed. Second, if urgency increases more workers could solve queued tasks more quickly. Third, the effect on the task acceptance rate also influences the number of workers. On the one hand, if queueing becomes more costly less tasks from customers will be accepted and the number of workers decreases. On the other hand, since the knowledge of workers increases, less tasks will be received from direct subordinates and so the task acceptance rate increases and the number of workers increases. We show that if the marginal revenue with respect to the expected number of solved tasks is elastic, it is optimal to increase the knowledge of the workers and to decrease the span of control of the knowledge manager. This implies that decisions are more decentralized, since the probability to forward a task to the superior knowledge manager decreases.

In a numerical example we analyze a specific change of the organizational structure, namely the flattening of the organization. In general it refers to the elimination of divisions in a firm. We consider three key players in the creation of knowledge in the organization, the junior manager, the middle manager and the senior manager and discuss the transformation or elimination of middle management. Colombo and Grilli (2013) show in an empirical study of Italian high-tech entrepreneurial ventures that the information overload problems are key-drivers for the creation of middle management. Consistent with this, we find that an overload problem for senior managers occurs when middle management is transformed to junior management. Wulf (2012) finds that after flattening, which should push decisions downwards, there is more control and decision making at the top of the organization. In our example, if middle management is eliminated or transformed into junior management, the task acceptance rate of the senior manager declines. His division has to solve comparably more forwarded tasks from subordinate knowledge managers and is therefore more involved in internal tasks. The elimination or transformation of the middle management results in a less hierarchical structure. A rational for flattening is that a streamlined firm may respond more quickly to customers. However, we find that the mean response time, i.e. the time a task spends

in the organization, is higher in the flattened organizations. The reason is that in the organization with a middle management, the task acceptance rate of the senior manager is higher. External tasks handled directly by the senior manager division are not forwarded to any other manager and therefore are served faster. Given the optimal organizational structure, we characterize the optimal wage contracts and find that for all managers the wage sensitivity to performance is higher under joint performance evaluation than independent performance evaluation. We consider two different types of joint performance evaluations. First, the bonus payment depends also on the output of the subordinate knowledge manager and second the bonus payment depends on the output of all knowledge managers. We find that while for the first type of performance evaluation a bonus contract is optimal, under the second type a penalty contract is optimal for the junior and middle managers. Wulf (2007) reports that for division managers of lower rank, the bonus payment is typically linked to performance measures over which the manager has greater control than the overall performance of the organization. This means for the numerical example that the first type of performance evaluation is more relevant to junior and middle managers.

2 The model

The model consists of: (1) an organizational component, (2) an informational component, and (3) an economic component. The structure of the organization incorporates besides the supervisory structure and the production process. The informational component specifies the information asymmetry between HRM and manager. The economic component comprises the objectives of HRM and the incentives for the managers.

2.1 Organizational structure

The organization has L divisions, where L is taken as given. A division is an organizational unit which consists of one knowledge manager M_l (the agent) and r_l workers. Let t_l be the arrival rate of tasks given by customers for division l, which means that on average t_l tasks arrive per unit of time. The knowledge manager M_l, to which we will refer to as the manager, coordinates the arriving tasks and forwards tasks which were processed but unsolved to managers of other divisions. The job of the workers in division l is to solve tasks given by their manager M_l. The

manager is in charge for the training of the workers in his division. We will therefore refer to r_l as the *span of control* of M_l is r_l. If the manager invests an effort of $p_l \in [0,1]$ in the training of his workers, they are able to solve a fraction p_l of the arriving tasks. HRM (the principal) is in charge of designing the payment scheme w_l of the manager M_l, decides how many workers are employed r_l in the division, and determines the task acceptance rate t_l of each manager M_l.

Corporate governance structure. There are two different types of supervision. First, each manager is supervising a division of r_l workers. Second, all managers, except the head of the organization directly report to a manager of higher rank. Tasks which where processed but could not be solved are forwarded by the manager to his immediate supervisor manager. We follow Cho (2010) who describes the second supervisory structure by a directed graph. The nodes are divisions and edges are links between divisions. Let $\bar{L} = \{1, \ldots, L\}$ be the set of the divisions. A *link between node l and k* is denoted by lk and is an ordered pair $(l,k) \in \bar{L} \times \bar{L}$. If the link kl exists, then M_l is a supervisor of M_k, i.e. M_k is a subordinate of M_l. A *path from node k to l* is $\{l_1 l_2, \ldots, l_{K-1} l_K\}$ where $l_1 = k$ and $l_K = l$. M_l *controls* M_k if there is a path from node k to l. Let C_l be the set of divisions that M_l controls and C_l^d be the set of divisions that M_l controls directly. If $k \in C_l$ and the link kl exists, then M_l *controls directly* M_k. If such a link does not exist, but $k \in C_l$ then M_l *controls indirectly* M_k. Division l is the *head division* if M_l is not controlled by any other division, i.e. $C_l = \bar{L} \setminus \{l\}$. Division l is a *low level division* if M_l is not supervisor to any other division, i.e. $C_l = \emptyset$. The structure of the organization is assumed to have the following properties

1. There is a unique head division.

2. Each division, except the head division has only one supervisor.

3. If manager M_l supervises M_k, then M_k cannot supervise M_l directly or indirectly.

The supervisory structure implies a rank of the division. We consider the counting up rank system (Beckmann 1988). The manager of a low level division has rank 0. The manager supervising only the manager of the low level division has rank 2 and so on. For a more formal definition and examples of the implied rank system see Appendix A.

Production process. Time is continuous and at each instance consumers send tasks to the divisions. We assume that the manager M_l chooses the arrival rate of tasks t_l, given the requirement

of HRM. More formally, suppose that potential tasks arrive at each division with a given rate T. M_l accepts the potential task with probability λ_l and rejects with probability $1-\lambda_l$. This decision is independent for successive customers and of the number of tasks in the system. By choosing λ_l, M_l makes sure that tasks arrive at each node according to independent Poisson processes with rate $t_l = \lambda_l T$ to which we will refer as the *task acceptance rate of* M_l. Since each division has a finite number of workers tasks might not be immediately dealt with. If all workers are busy the incoming tasks are queued. M_l distributes the queued tasks evenly to the queues of the workers in his span of control. The queue at each worker can be of infinite length and tasks are processed on a first-come-first-serve basis and the service time follows an exponential distribution. For simplicity assume that on average, if the queue is busy, one task can be handled by each worker per unit of time, i.e. the service rate at each server is $\mu_l = 1$. Independent of other tasks, each arriving task is associated with a difficulty level P which follows a uniform distribution over $[0,1]$. The workers in each division have to acquire a certain knowledge measured by $p_l \in [0,1]$, which is the probability that a task is solved. The output of division l, i_l, is the number of solved tasks of division l. Suppose the link kl exists. If workers in division k cannot solve a task, which happens with probability $1-p_k$, the manager M_k forwards the task to its immediate supervisor M_l who forwards the task to an idle worker or queues it. The probability to forward a processed task to another division is independent of its past history and independent of all other tasks. The supervisory structure determines if and to whom unsolved tasks can be passed on. In queueing networks these relationships are captured by the balance equations which state that in the long-run the rate of flow of tasks out of a division has to equal the rate of flow of tasks into the division.

$$a_l = t_l + \sum_{j=1}^{L} a_j p_{jl} \qquad l = 1, \ldots, L$$

where a_l is the effective arrival rate at division l to which we will refer as the *task arrival rate of* M_l, and p_{jl} is the proportion of unsolved tasks passed on from division j to division l with $p_{ll} = 0$, i.e. unsolved tasks are only passed on to other layers. Since we consider hierarchical organizations, where division j is allowed to pass only to one other division l so $p_{jl} = 1 - p_j$ if the link jl exits

and $p_{jl} = 0$ otherwise. The balance equations in matrix notation are

$$a = t + \overline{P}a \tag{1}$$

where $\overline{P}$ is the routing matrix and $\overline{P}_{ij}$ gives the proportion of tasks forwarded from division j to division i. Since the communication structure is uni-directional $\overline{P}$ will be lower triangular, with $\overline{P}_{ii} = 0$. Equation 1 can be solved for a

$$a = Pt \tag{BE}$$

where $P := \left(I - \overline{P}\right)^{-1}$ and I is the identity matrix. Properties of P are discussed in appendix B. Since P is also lower triangular, the balance equation for division l can be written as

$$a_l = \sum_{k=1}^{L} P_{lk} t_k = t_l + \sum_{k:k \in C_l} P_{lk} t_k \tag{2}$$

In order for the queueing problem to be well defined it has to hold that the net inflow of each division is less than the total service rate

$$\rho_l = \frac{a_l}{r_l} < 1$$

where ρ_l is the *traffic intensity of division* l. It is mean number of tasks in service at a worker in division l and it is equal to the percentage of time a worker is busy. For such queueing networks, Jackson (1963) showed that the joint probability distribution for the number of tasks at each division is the product of the marginal probability distributions at each division. So the performance of each division can be analysed independently as a $M/M/1$ queue for each worker, i.e. a single queue at each worker where external arriving tasks follow a Poisson distribution and task service time follows an exponential distribution.

An example is given in Figure 1. There are two divisions, where division 1 has four workers, $r_1 = 4$, and division 2 has two workers, $r_2 = 2$. The manager of division 2, M_2, is supervisor of the manager of division 1, M_1. M_1 has task acceptance rate of t_1 and M_2 has a task acceptance rate

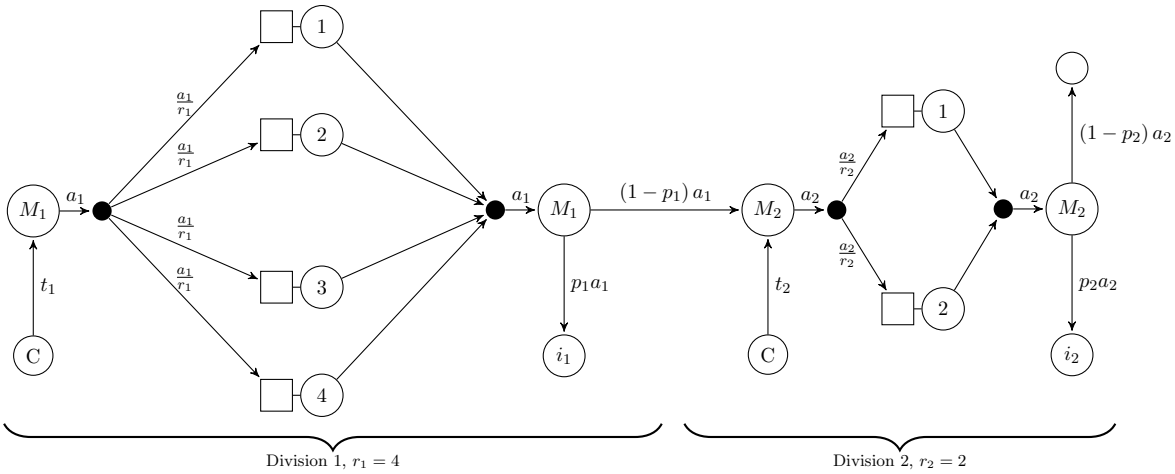

Figure 1: A queuing network with 4 workers (queues) in division 1 and 2 workers (queues) in division 2. For $l = 1, 2$, M_l is the manager from division l. Each division solves on average $p_l a_l$ tasks per unit of time. M_2 accepts t_2 tasks from clients and $(1 - p_1)\, a_1$ tasks from the subordinate manager M_1. On average $(1 - p_2)\, a_2$ tasks remain unsolved.

of t_2 and receives $(1 - p_1)\, a_1$ tasks from M_1. The balance equations are

$$a_1 = t_1$$

$$a_2 = t_2 + (1 - p_1)\, a_1$$

Each worker in division 1 receives task with rate $\frac{a_1}{4}$. After the workers have dealt with the task M_1 separates the tasks into solved cases, the output of division 1, i_1, and unsolved cases which are forwarded to M_2. On average division 1 solves $p_1 a_1$ cases. Each worker in division 2 receives tasks with rate $\frac{a_2}{2}$. The average output of division 2 is $p_2 a_2$ and $(1 - p_2)\, a_2$ tasks remain unsolved.

2.2 Information Structure

The effort choice p_l of M_l influences the probability of solved tasks. HRM cannot observe this effort choice but only the number of solved tasks. HRM designs the wage contracts for the managers of the divisions, w_l, which depends on the number of solved tasks i.

2.3 Economic Structure

Performance Measures. Suppose that the organization creates revenue by solving tasks. Thus, one possible performance measure is the average number of solved tasks of each division.

Suppose that for each completed task the firm receives a gain of H_l. For the network of queues considered it holds that under the equilibrium distribution, the external departure of tasks also follow independent Poisson processes (Jackson 1963). So the external departure (output) follows also a Poisson process with rate $\theta_l = p_l a_l$. The average number of solved tasks of each division is θ_l to which we will refer as the *throughput of division* l. Suppose that HRM evaluates the throughput of each division by a revenue function $H_l(\theta_l)$, which is assumed to be strictly increasing and strictly concave, $H_\theta > 0$ and $H_{\theta\theta} < 0$. H_θ denotes the first derivative and $H_{\theta\theta}$ the second derivative with respect to θ_l. Since divisions with a higher rank might charge more for their solved problems, H_l is assumed to be nondecreasing in l.

Another performance measure is the mean number of tasks pending in the organization. Tasks might not be immediately dealt with and are queued. If it is costly to have tasks queued, it is in the interest of the organization to keep the queues short. Reasons why it is costly for the organization could be storage costs or just it does not want to keep customers waiting too long for service. The mean number of tasks pending in the division l is the mean number of total tasks in division l minus the tasks in service

$$Q = \sum_{l=1}^{L} Q_l = \sum_{l=1}^{L} \frac{a_l^2}{r_l - a_l}$$

Cost Structure. The workers in division l receive a fixed wage of $c_l^W > 0$. We assume that workers in a division of higher rank have a higher fixed wage, i.e. if M_l has a higher rank than M_k then $c_l^W > c_k^W$. The managers are assumed to be risk neutral and receive payment scheme, w_l, designed by HRM. Additionally, the manager has a utility loss due to knowledge management. We assume that the loss is additive in the effort of the manager of creating knowledge and the number of workers in his division. If the manager invests an effort of $p_l \in (0,1)$ in knowledge creation and training of the workers in his division, they are able to solve a fraction of p_l of the arriving tasks and he has a disutiliy of $G_l(p_l) = g_l(p_l) + F_l$. G_l is increasing in p_l and strictly convex , $g_p > 0$ and $g_{pp} > 0$. g_p denotes the first derivative and g_{pp} the second derivative with respect to p_l. If there are r_l workers in his span of control an additional loss of $c_l^M r_l$ occurs, with $c_l^M > 0$.

Objective of HRM. The objective of HRM is to maximize a function comprised of the key

performance indicators for each division

$$K_l = H_l(\theta_l) - \beta \frac{a_l^2}{r_l - a_l} - \mathbb{E}[w_l(I)|p] - c_l^W r_l$$

where the random variable $I = (I_1, \ldots, I_L)$ is the output of all divisions. Each division generates costs, the wage of the manager w_l, the wage for each worker c_l^W, and queueing costs βQ_l. HRM maximizes the sum of profits of each division. It is assumed that the average number of tasks in the queue affects the organization value linearly by a factor of $\beta \in (0,1)$. If β is high, then a long queue is more costly for the organization.

2.4 Problem Formulation

HRM considers the following maximization problem

$$\max_{w,t,p,r} \sum_{l=1}^{L} K_l$$

$$s.t.\ \mathbb{E}[w_l(I)|t] - G_l(p_l) - c_l^M r_l \geq 0 \text{ for } l = 1, ..., L \quad \text{(IR)}$$

$$p_l \in \arg\max_{p_l'} \left\{ \mathbb{E}\left[w_l(I)|p'\right] - G_l(p_l) - c_l^M r_l \right\} \text{ for } l = 1, ..., L \quad \text{(IC)}$$

$$w_l(i) \geq 0 \text{ for } l = 1, \ldots, L \text{ and } \forall i \quad \text{(LL)}$$

$$a = Pt \quad \text{(BE)}$$

$$p_l \in [0,1] \text{ for } l = 1, \ldots, L$$

$$a_l < r_l \text{ for } l = 1, \ldots, L$$

$$r_l \geq 0 \text{ for } l = 1, \ldots, L$$

where $i = (i_1, \ldots, i_L)$ is the observed output of all divisions.

In section 3 the optimal organizational structure is determined, when the effort of the manager is contractable. In that case HRM offers a wage w_l such that it is individual rational for M_l to accept the contract, i.e. condition (IR) is binding. Since HRM can observe the effort level condition it does not need to take into account the incentive compatible condition (IC). We characterize the optimal task acceptance rate t_l^*, the optimal proportion of solved tasks p_l^* and the optimal number of workers in each division r_l^*. In section 4, under the assumption that the effort of the manager

is not observable, a bonus contract is characterized. If the effort of the manager is not observable, HRM has to take condition (IC) into account as well. It is shown that under some conditions the managers M_l will accept the contract and the effort level under full information can be implemented.

3 Analysis: The Organizational Structure

In this section the optimal organizational form is determined under the assumption that the effort of the manager is contractable. In that case it is optimal for HRM to offer the manager an expected wage equal to his reservation utility. Since p_l is contractable, HRM does not need to take the incentives of the manager into account. This means that condition IR is binding and condition IC can be omitted. HRM solves the following simplified optimization problem:

$$\begin{aligned} &\max_{t,p,r} \sum_{l=1}^{L} \left(H_l(\theta_l) - \beta \frac{a_l^2}{r_l - a_l} - G_l(p_l) - c_l^M r_l - c_l^W r_l \right) \\ &s.t.\ a = Pt \qquad \text{(BE)} \\ &\quad p_l \in [0,1] \text{ for } l = 1, \ldots, L \\ &\quad r_l \geq 0 \text{ and } a_l < r_l \text{ for } l = 1, \ldots, L \\ &\quad t_l \geq 0 \text{ for } l = 1, \ldots, L \end{aligned}$$

The following first order conditions are derived in Appendix E.1. The first order conditions for the optimal task acceptance rate t_l^* is

$$H_\theta(\theta_l)\, p_l = \beta \frac{2 a_l r_l - (a_l)^2}{(r_l - a_l)^2} \text{ for } l = 1, \ldots, L-1 \tag{3}$$

At the optimum it has to hold that the marginal gain from solving a task has to be equal to the marginal loss of an additional task pending at division l. The right hand side is positive since at the optimum it holds that $r_l > a_l$. The marginal effect on other divisions, $\frac{\partial a_k}{\partial t_l}$, does not play a role due to the hierarchical structure of the organization. The size of t_l however, does play a role on other divisions through $a_k^* = \sum_{m=1}^{l} P_{km} t_m^*$.

The first order conditions for p_l^* is

$$H_\theta(\theta_l)\, a_l = g_p(p_l) \tag{4}$$

The marginal gain of additional knowledge in division l has to be equal to the marginal cost of the manager M_l of coaching the workers in his span of control to attain that knowledge.

The first order conditions for r_l^* is

$$\beta \frac{a_l^2}{(r_l - a_l)^2} = c_l$$

where $c_l = c_l^W + c_l^M$. The marginal gain of an additional worker by reducing the task pending at division l has to be equal to the marginal cost of paying one additional worker. It can be simplified to

$$r_l^* = a_l^* \frac{\sqrt{c_l} + \sqrt{\beta}}{\sqrt{c_l}} \tag{5}$$

so $r_l^* > a_l^*$ will hold.

The traffic intensity of each division is

$$\rho_l^* = \frac{a_l^*}{r_l^*} = \frac{\sqrt{c_l}}{\sqrt{c_l} + \sqrt{\beta}} \tag{6}$$

and the variable cost of knowledge creation per worker are

$$\frac{p_l^* g_p(p_l^*)}{r_l^*} + c_l^M = \frac{c_l^{3/2} + 2c_l\sqrt{\beta}}{\sqrt{c_l} + \sqrt{\beta}} + c_l^M$$

Under the assumption that c_l^W and c_l^M are increasing in rank, traffic intensity and the variable cost of knowledge management per worker are increasing in rank of the manager. This means that on average a worker of a division of higher rank has more tasks in service than a worker of a division of lower rank. A manger of higher rank has higher variable cost of knowledge management than a manager of lower rank.

The optimal number of external tasks t_l can be determined from equation BE

$$t^* = P^{-1}a = \left(I - \bar{P}\right) a^*$$

Due to the hierarchical structure, the task acceptance rate of division l depends only on divisions directly controlled by M_l.

$$t_l^* = a_l^* - \sum_{k:k \in C_l^d} \left(1 - p_k^*\right) a_k^* \tag{7}$$

Suppose M_l controls only M_k directly and M_k controls only M_m directly, then

$$t_l^* < t_k^* \leftrightarrow a_l^* - a_k^* < \left(1 - p_k^*\right) a_k^* - \left(1 - p_m^*\right) a_m^*$$

M_l will have lower task acceptance rate than M_k if the difference of the task arrival rate is smaller than the difference of the arriving fraction of unsolved problems. This means that M_l receives comparably more unsolved problems and therefore has a lower task acceptance rate.

In appendix E the second order conditions are derived. In order for a critical point to be a maximum it has to hold that $-{}^{H_{\theta\theta}(\theta_l)\theta_l}/_{H_\theta(\theta_l)} > 0.5$, which means that the absolute value of the elasticity of the marginal gain of throughput is higher than 0.5.

3.1 Comparative Static

The parameter β measures the intensity of the queueing costs and c_l are the total cost per worker. In this section the effect of β and c_l on p_l^*, r_l^*, t_l^* and ρ_l^* are discussed. Combining equation 3, 4 and 5 gives two equations in p_l^* and r_l^*

$$H_\theta\left(\theta_l^*\right) p_l^* = c_l + 2\sqrt{\beta\left(c_l\right)} \tag{8}$$

$$H_\theta\left(\theta_l^*\right) a_l^* = g_p\left(p_l^*\right) \tag{9}$$

with $a_l^* = \frac{r_l^* \sqrt{c_l}}{\sqrt{c_l} + \sqrt{\beta}}$. In order to derive the effects we apply the implicit function theorem on equations 8 and 9. The derivations are given in Appendix E.2.

Suppose that at the optimum the elasticity of the marginal gain of throughput is higher than

unity

$$1 < -\frac{H_{\theta\theta}\left(\theta_l^*\right)\theta_l^*}{H_\theta\left(\theta_l^*\right)}$$

Then it can be shown that

$$\frac{\partial p_l^*}{\partial \beta} \geq 0, \frac{\partial p_l^*}{\partial c_l} \geq 0, \frac{\partial r_l^*}{\partial c_l} \leq 0, \frac{\partial r_l^*}{\partial \beta} \leq 0$$

If the elasticity of the marginal gain of throughput is equal to 1, then a change in β or c_l has no effect on p_l. If $H_{\theta\theta}a_l^2 = g_{pp}$, a change in c_l has no effect on r_l.

If urgency increases, i.e. it becomes more important that tasks are not pending, the manager increases training of the workers in his span of control. This implies that the probability to solve a task directly at each division increases. There are three effects on the number of workers in each division. First, since workers are able to solve more problems, less workers will be employed. Second, if urgency increases more workers could solve queued tasks more quickly. Third, the effect on the task acceptance rate also influences the number of workers. Under the assumption that the marginal gain of throughput is elastic the span of control will decrease. Otherwise it may occur that higher urgency results in more workers. If the wage costs of the workers increases again it is optimal for the manager to train the workers more. Since workers are more costly and are able to solve more problems less workers are employed. This implies that in both cases knowledge increases and less tasks are forwarded to superiors, decisions are more decentralized but divided under less workers.

Although the effect on p_l and r_l depends on the assumption on elasticities of marginal gain, the effect on traffic intensity is always unambiguous. From equation 6

$$\frac{\partial \rho_l^*}{\partial \beta} = -\frac{\sqrt{c_l}}{2\sqrt{\beta}\left(\sqrt{c_l}+\sqrt{\beta}\right)^2} < 0$$
$$\frac{\partial \rho_l^*}{\partial c_l} = \frac{\sqrt{\beta}}{2\sqrt{c_l}\left(\sqrt{c_l}+\sqrt{\beta}\right)^2} > 0$$

If urgency increases traffic intensity will decrease. It means, that on average a worker has less tasks in service. Since in this model it holds that $\rho_l = \frac{Q_l}{N_l}$ (see Appendix D) it also means that the ratio of

tasks pendiong to total tasks in the division is decreasing. If the wage cost of the workers increase, traffic intensity will increase.

From equation 7 the effect of urgency on the task acceptance rate can be determined

$$\frac{\partial t_l^*}{\partial \beta} = \frac{\partial r_l^*}{\partial \beta}\frac{\sqrt{c}_l}{\sqrt{c}_l + \sqrt{\beta}} + \frac{\partial a_l^*}{\partial \beta} + \sum_{k:k\in C_l^d} \frac{\partial p_k^*}{\partial \beta} a_k - (1-p_k)\left(\frac{\partial r_k^*}{\partial \beta}\frac{\sqrt{c}_k}{\sqrt{c}_k + \sqrt{\beta}} + \frac{\partial a_k^*}{\partial \beta}\right)$$

The first term, which captures the effect of an increase in urgency on the own task arrival rate, is negative. The second term, which captures the effect on unsolved problems from all other divisions forwarded to M_l, is positive. This effect is positive since the knowledge in divisions increases and less tasks are forwarded. Therefore more tasks can be accepted from customers. If the decrease in own task arrival rate outweighs the decrease in unsolved problems then the task acceptance rate has to decrease as well. If the decrease of unsolved problems in other divisions predominates, then task acceptance rate will be higher. The effect of an increase in total wage costs of the worker is also ambiguous

$$\frac{\partial t_l^*}{\partial c_l} = \frac{\partial r_l^*}{\partial c_l}\frac{\sqrt{c}_l}{\sqrt{c}_l + \sqrt{\beta}} + \frac{\partial a_l^*}{\partial c_l}$$

The first term is negative while the second term is positive. The effect of an increase in wage cost of the workers on the task acceptance rate does not depend on the other divisions. Higher wage costs of workers decrease the number of workers so the task has to be distributed to fewer workers. This has a negative effect on task acceptance rate. On the other hand, higher wage cost increase the knowledge of the workers which has a positive on the task arrival rate and so the task acceptance rate can be increased. Suppose that only in division l the total wage costs of workers increase, then the effect on the task acceptance rate of another division k is zero if l is not a direct subordinate of k and otherwise

$$\frac{\partial t_k^*}{\partial c_l} = \frac{\partial p_l^*}{\partial c_l} a_l^* - (1-p_l)\left(\frac{\partial r_l^*}{\partial c_l}\frac{\sqrt{c}_l}{\sqrt{c}_l + \sqrt{\beta}} + \frac{\partial a_l^*}{\partial c_l}\right) = \frac{\partial p_l^*}{\partial c_l} a_l^* - (1-p_l)\frac{\partial t_l^*}{\partial c_l}$$

So if the task acceptance rate in division l decreases then it will increase in division k.

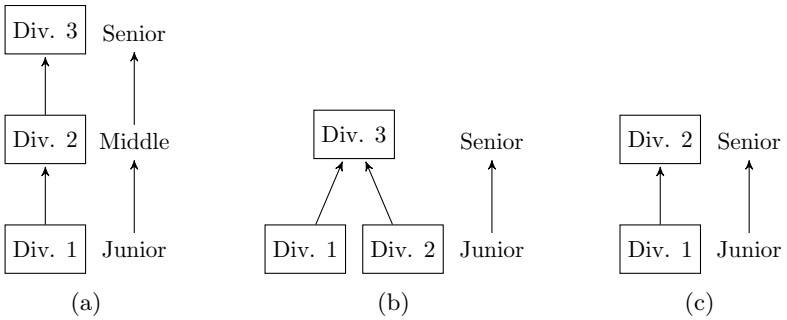

Figure 2: Organization A consists of three divisions, where the manager of each division has a different rank (Figure 2(a)). Organization B consists of three divisions, where division one and two are guided by junior managers and division three is guided by a senior manager (Figure 2(b)). Organization C has two divisions led by managers of different rank (Figure 2(c)).

3.2 Numerical Example

In the numerical example we discuss the effects of the elimination of the middle management division. We consider three different organizational structures. Organization A (Figure 2(a)) has one junior manager (division 1), one middle manager (division 2) and one senior manager (division 3). In Organization B (Figure 2(b)) the middle management division is transformed into a junior management division. In Organization C (Figure 2(c)) the middle management division is eliminated. The assumed specific functional forms for the disutilities of the manager G_l, the revenue of each division H_l, and the wage cost of the workers c_l^W are given in Table 1. For simplicity, we set the wage costs of the workers to zero so that the total costs of workers are the disutility a manager has from coaching, $c_l = c_l^M$. Since the optimal number of workers in each division depends only on the total costs c_l, the same results would be obtained if a share of c_l is interpreted as the wage costs. In order to make organization A and C comparable we chose the parameters such that for $\beta = 0.8$, their profits are equal, $K_l^A = K_l^C = 0$.

From the first order conditions it follows that for division l with rank s

$$p_{ls} = \sqrt{\frac{h_{ls}}{x_{ls}}}$$

$$a_{ls} = \frac{h_{ls}}{c_{ls} + 2\sqrt{(c_{ls})\,\beta}}$$

$$r_{ls} = a_{ls}\left(1 + \sqrt{\frac{\beta}{c_{ls}}}\right)$$

Table 1: Functional forms and parameter

Knowledge management cost of manager M_l	
due to coaching	$G_l(p_l) = g_l(p_l) + F_l$
	$g_l(p_l) = \frac{x_l}{2}(p_l)^2$
due to span of control	$c_l^M r_l$
Revenue of division l	$H_l(\theta_l) = h_l \ln(\theta_l)$
Intensity of queueing costs	$\beta \in \{0.5, 0.8\}$
Wage costs of workers	$c_l^W = 0$

	Junior	Middle	Senior
h_l	28.09	35.22	36.80
c_l^M	0.37	0.77	1
x_l	70.22	44.02	36.80
F_l	28.09	38.74	39.75

At the optimum the gain of HRM is

$$K_{ls} = H_{ls}(\theta_{ls}) - \beta \frac{a_{ls}^2}{r_{ls} - a_{ls}} - G_{ls}(p_{ls}) - R_{ls}(r_{ls}) = h_{ls}\left(\ln\left(\frac{(h_{ls})^{\frac{3}{2}}}{\sqrt{x}_{ls}(c_{ls} + 2\sqrt{c}_{ls}\beta)}\right) - \frac{3}{2}\right)$$

In the simple example, the optimal amount of coaching p^{ls} is independent of β and c_l since $-H_{\theta\theta}\theta_l/H_\theta = 1$. The optimal task acceptance rate can be derived from the balance equation

$$t^s = \left(I - \overline{P}^s\right) a^s$$

where

$$\overline{P}^A = \begin{bmatrix} 0 & 0 & 0 \\ 1-p_1 & 0 & 0 \\ 0 & 1-p_2 & 0 \end{bmatrix}, \overline{P}^B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1-p_1 & 1-p_2 & 0 \end{bmatrix}, \overline{P}^C = \begin{bmatrix} 0 & 0 \\ 1-p_1 & 0 \end{bmatrix}$$

Let $\beta = 0.8$. The results are given in Table 2 and 3. In the example for all three organizations, the span of control of the managers r_l and the task arrival rate a_l are decreasing in the rank of the manager. However, the traffic intensity ρ_l is increasing. This implies that an worker in a division of higher rank is busy at a higher percentage of time than an worker in a division of lower rank. The effort of the manager in coaching p_l is increasing in rank which implies that a division of higher rank solves on average more tasks than a division of lower rank. The task acceptance rate can be non monotonic in the rank of the managers. In organization A, the senior manager has a higher task acceptance rate than the middle manager. The middle manager receives comparably more unsolved tasks than the senior manager and therefore can accept less tasks from customers in order

Table 2: Results for organization A,B and C for $\beta = 0.8$

Rank	A,B,C Junior	A Middle	A Senior	B Senior	C Senior
Effort in coaching, p_l	0.63	0.89	1	1	1
Number of Workers, r_l	49.37	31.63	26.03	28.65	26.03
Effective arrival rate of tasks, a_l	22.83	17.51	15.25	16.78	15.25
Task acceptance rate, t_l	22.83	9.12	13.40	0	6.85
Traffic intensity, ρ_l	0.46	0.55	0.59	0.59	0.59
Throughput, θ_l	14.44	15.66	15.25	16.78	15.25
Tasks pending, Q_l	19.64	21.73	21.56	23.73	21.56
Tasks pending and in service, N_l	42.47	39.25	36.80	40.51	36.80
Variable costs of knowledge management, $g_l(p_l)$	14.04	17.61	18.40	18.40	18.40
Costs due to span of control, $c_l^M r_l$	18.27	24.35	26.03	28.65	26.03
Queuing costs, βQ_l	9.82	10.87	10.78	11.87	10.78
Revenue, $H_l(\theta_l)$	74.99	96.90	100.27	103.80	100.27
Profit, K_l	4.77	5.33	5.31	5.13	5.31

for the balance equation to be fulfilled. Also the mean number of tasks pending in each division is non monotonic in rank. The average queue length of the senior manager is less than of the middle manager. So the main job of middle manager are solving tasks given by customers and relieving the senior manager from a potential overflow of unsolved tasks from the junior manager.

Eliminating the middle management can result in organization B or C. If instead of one middle manager two junior manager report to the senior manager (Figure 2(b)) the task arrival rate of the senior manager increases and in order for the balance equation to be fulfilled, it would be required that $t_2 < 0$. This means, that there is an overflow of tasks to the head division. If the division of the middle manager is eliminated without replacement (Figure 2(c)) then the problem of overflow does not occur. However, the task acceptance rate of the senior manager decreases since unsolved tasks are not handled by middle managers first.

With respect to the overall gain, HRM is indifferent between organization A or C. However, these two organizations differ in their structure. Although organization C is flatter and has on average less tasks pending, the mean response time W, i.e. the time a task spends in the organization, is higher in organization C than in A. The formal definition is given in Appendix D. The reason is that in organization A the task acceptance rate of the senior manager is higher and if a task is directly handled by the senior manager it is served faster.

Suppose that a delay becomes less costly, $\beta = 0.5$. The results are given in Table 4 and 5.

Table 3: Results for organization A,B and C for $\beta = 0.8$

Organization	A	B	C
Total number of workers, $\sum_{l=1}^{L} r_l$	107.02	127.39	75.40
Total effective arrival rate of tasks, $\sum_{l=1}^{L} a_l$	55.59	62.44	38.08
Total task acceptance rate, $\sum_{l=1}^{L} t_l$	45.35	45.66	29.68
Total tasks pending, $\sum_{l=1}^{L} Q_l$	62.93	63.01	41.20
Total tasks pending and in service,$\sum_{l=1}^{L} N_l$	118.52	125.46	79.27
Q/N	0.53	0.50	0.52
Total throughput, $\sum_{l=1}^{L} \theta_l$	45.35	45.66	29.68
Response time, W	2.61	2.75	2.67
Total variable costs of knowledge management, $\sum_{l=1}^{L} G_l(p_l)$	156.63	142.41	100.28
Total costs due to span of control $\sum_{l=1}^{L} R_l(r_l)$	68.64	65.18	44.29
Total queueing costs, $\sum_{l=1}^{L} \beta Q_l$	31.47	31.51	20.60
Total revenue, $\sum_{l=1}^{L} H_l(\theta_l)$	272.16	253.78	175.26
Total profit, $\sum_{l=1}^{L} K_l$	15.42	14.68	10.08

Although in this example the decrease in urgency has no effect on the knowledge of the workers, it results in a higher span of control. This means that the effect of a higher task acceptance rate, which results in more workers, outweighs that a decrease in urgency could imply fewer workers. In organization A, the division of the middle manager has the highest gain, which also implies that by removing division 2 or transforming it into a division of rank 0 does not result in higher total gain.

4 Analysis: The Incentive Structure

Under asymmetric information HRM cannot observe the effort a manager puts into coaching of the workers in his span of control, but the amount of solved tasks of each division. We consider two different types of performance evaluation with bonus scheme: first, under independent performance evaluation a bonus is paid, if the division reached the targeted output and second, under joint performance evaluation a bonus is paid only if all divisions have achieved their target. We identify conditions, which depends crucially on the hazard rate of effort, under which such a bonus contract exist. The conditions are similar to Kim (1997) and Park (1995) who considered similar bonus contracts for a single agent under asymmetric information and limited liability. We follow their lines of argument and extend a bonus contract for the multiple agent case. The hazard rate of effort gives the change in the cumulative distribution of output resulting from higher effort in relation to

Table 4: Results for organization A,B and C for $\beta = 0.5$

Rank	A,B,C Junior	A Middle	A Senior	B Senior	C Senior
Effort in coaching, p_l	0.63	0.89	1	1	1
Number of Workers, r_l	47.59	30.40	25.00	26.82	25.00
Effective arrival rat of tasks, a_l	19.26	15.05	13.20	14.16	13.20
Task acceptance rate, t_l	19.26	7.97	11.61	0	6.12
Traffic intensity, ρ_l	0.40	0.50	0.53	0.53	0.53
Throughput, θ_l	12.18	13.46	13.20	14.16	13.20
Tasks pending, Q_l	13.10	14.77	14.75	15.83	14.75
Tasks pending and in service, N_l	32.36	29.82	27.95	30.00	27.95
Variable costs of knowledge management, $g_l(p_l)$	14.04	17.61	18.40	18.40	18.40
Costs due to span of control, $c_l^M r_l$	17.61	23.41	25.00	26.82	25.00
Queuing costs, βQ_l	10.48	11.81	11.80	12.66	11.80
Revenue, $H_l(\theta_l)$	70.22	91.57	94.96	97.55	94.96
Profit, K_l	0	0	0	-0.09	0

the probability to receive a bonus.

4.1 Independent Performance Evaluation

Under independent performance evaluation (IPE) the payment of the bonus only depends on the output of each division i_l'.

Proposition 1 (IPE)**.** *If there exists some level of solved tasks* $0 < i_l' \leq \lceil \theta_l \rceil - 1$ *such that*

$$\frac{g_p(p_l^*)}{G_l(p_l^*)} > -\frac{\frac{\partial F_l(i_l', \theta_l^*)}{\partial p_l}}{(1 - F_l(i_l', \theta_l^*))} > \frac{g_p(p_l^*)}{G_l(p_l^*) + c_l^M r_l^*} \tag{10}$$

where $F_l(i_l', \theta_l)$ *is the cumulative distribution of the individual success probability*

$$F_l(i_l', \theta_l) = \sum_{i_l=0}^{i_l'} \pi(i_l, \theta_l)$$
$$\frac{\partial F_l(i_l', \theta_l)}{\partial p_l} = -a_l \pi(i_l', \theta_l) < 0$$

then the following incentive scheme induces the first-best effort level p_l^* *and the manager receives*

Table 5: Results for organization A,B and C for $\beta = 0.5$

Organization	A	B	C
Total number of workers, $\sum_{l=1}^{L} r_l$	102.98	122.00	72.59
Total effective arrival rate of tasks, $\sum_{l=1}^{L} a_l$	47.51	52.68	32.46
Total task acceptance rate, $\sum_{l=1}^{L} t_l$	38.84	38.52	25.38
Total tasks pending, $\sum_{l=1}^{L} Q_l$	42.62	42.03	27.85
Total tasks pending and in service,$\sum_{l=1}^{L} N_l$	90.13	94.71	60.31
Q/N	0.47	0.44	0.46
Total throughput, $\sum_{l=1}^{L} \theta_l$	38.84	38.52	25.38
Response time, W	2.32	2.46	2.38
Total variable of knowledge management, $\sum_{l=1}^{L} G_l(p_l)$	156.63	142.41	100.28
Total costs due to span of control $\sum_{l=1}^{L} R_l(r_l)$	66.01	62.04	42.61
Total queueing costs, $\sum_{l=1}^{L} \beta Q_l$	34.10	33.62	22.28
Total revenue, $\sum_{l=1}^{L} H_l(\theta_l)$	256.74	237.98	165.17
Total profit, $\sum_{l=1}^{L} K_l$	0	-0.09	0

his reservation level of utility.

$$w_l(i_l) = \begin{cases} A_l + B_l & \text{for } i_l \geq i_l' \\ A_l & \text{else} \end{cases} \tag{11}$$

where $A_l = G_l(p_l^*) + c_l^M r_l^* - (1 - F_l(i_l', \theta_l^*)) B_l$ *and* $B_l = -\frac{g_p(p_l^*)}{\frac{\partial F_l(i_l', \theta_l^*)}{\partial p_l}}$

The proof and all following proofs are given in Appendix F.

The salary A_l and the bonus B_l depend only on the own output of each division. Under symmetric information it holds that $H_\theta(\theta_l^*) a_l^* = g_p(p_l^*)$, therefore

$$B_l = \frac{g_p(p_l^*)}{a_l^* \pi(i_l', \theta_l^*)} = \frac{H_\theta(\theta_l^*)}{\pi(i_l', \theta_l^*)}$$

The size of the bonus depends on the ratio of the marginal gain of one additional solved task and the probability that the target output is reached. Increasing the target output results in a lower bonus since $\pi(\theta_l, i_l')$ is increasing in i_l' for $i_l' \leq \lceil \theta_l \rceil - 1$ (See Appendix C, remark 1). The salary is

$$A_l = G_l(p_l^*) + c_l^M r_l^* - \left(1 - F_l\left(i_l', \theta_l^*\right)\right) B_l = G_l(p_l^*) + c_l^M r_l^* - H_\theta(\theta_l^*) \frac{(1 - F_l(i_l', \theta_l^*))}{\pi(i_l', \theta_l)} \tag{12}$$

Since the third term is decreasing in i_l', the salary will get closer to the expected wage $G_l(p_l^*) + c_l^M r_l^*$

when the target output is increased.

4.2 Joint Performance Evaluation

Under joint performance evaluation (JPE) the payment of the bonus depends on the output of all divisions, $i' = (i'_1, \ldots, i'_l)$. Proposition 2 characterizes the salary A_l, the bonus B_l, and i' the target output for which a bonus will be paid.

Proposition 2 (JPE)**.** *If there exists some level of solved tasks* $0 < i' \leq \lceil \theta \rceil - 1$ *such that*

$$-\frac{\frac{\partial F(i',\theta^*)}{\partial p_l}}{1 - F\left(i', \theta^*\right)} > \frac{g_p\left(p_l^*\right)}{G_l\left(p_l^*\right) + c_l^M r_l^*} \quad \text{if} \quad \frac{\partial F\left(i', \theta\right)}{\partial p_l} < 0 \tag{13}$$

$$\frac{\frac{\partial F(i',\theta^*)}{\partial p_l}}{F\left(i', \theta^*\right)} > \frac{g_p\left(p_l^*\right)}{G_l\left(p_l^*\right) + c_l^M r_l^*} \quad \text{if} \quad \frac{\partial F\left(i', \theta\right)}{\partial p_l} > 0 \tag{14}$$

$$\frac{F\left(i', \theta^*\right) - \lim_{p_l \to 0} F\left(i', \theta\right)}{\frac{\partial F(i',\theta^*)}{\partial p_l}} \geq \frac{g_p\left(p_l^*\right)}{G_l\left(p_l^*\right)} \tag{15}$$

where $F\left(i', \theta\right)$ *is the cumulative distribution of the joint success probability*

$$F\left(i', \theta\right) = \prod_{l=1}^{l} F_l\left(i'_l, \theta_l\right)$$

$$\frac{\partial F\left(i', \theta\right)}{\partial p_l} = -\sum_{k: l \in C_k} \frac{\partial \theta_k}{\partial p_l} \pi\left(i'_k, \theta_k\right) \prod_{m \neq k}^{L} F_m\left(i'_m, \theta_m\right) - a_l \pi\left(i'_l, \theta_l\right) \prod_{m \neq l}^{L} F_m\left(i'_m, \theta_m\right)$$

then the following incentive scheme induces the first-best effort level p_l^* *and the manager receives his reservation level of utility.*

$$w_l\left(i_l\right) = \begin{cases} A_l + B_l & \text{for } i_l \geq i'_l \\ A_l & \text{else} \end{cases} \tag{16}$$

where $A_l = G_l\left(p_l^*\right) + c_l^M r_l^* - \left(1 - F\left(i', \theta^*\right)\right) B_l$ *and* $B_l = -\frac{g_p\left(p_l^*\right)}{\frac{\partial F(i',\theta^*)}{\partial p_l}}$

The size of the bonus B_l depends on the effect of an increase in coaching on the own success probability and on the success probabilities of divisions which supervise division l.

4.3 Discussion of IPE and JPE

If target output is less or equal to the mode of the success probability, $\lceil \theta_l \rceil - 1$, it is sufficient for the success probability to be strictly convex at the optimum and therefore the maximization problem of the manager to be strictly concave at the optimum. In that case p_l^* will be a local maximum. This is however a sufficient but not necessary condition in order for the maximization problem of the manager to be strictly concave at the optimum. In the example we show that also for target output higher than the mode a penalty contract exists such that p_l^* maximizes the utility of the manager.

The conditions for a bonus contract to exist are stronger und JPE than under IPE.

Lemma 1. *If* $\partial F(i',\theta^*)/\partial p_l < 0$ *and if for some* $i' = (i'_1, \ldots, i'_l)$ *the condition for JPE is fulfilled, then it is also fulfilled at* i'_l *for IPE.*

This follows since $1 - F_l\left(i'_l, \theta_l^*\right) < 1 - F\left(i', \theta^*\right)$ and $-\frac{\partial F_l\left(i'_i, \theta_l^*\right)}{\partial p_l} > -\frac{\partial F(i', \theta^*)}{\partial p_l}$. It means that if a contract under JPE can be implemented, a contract under IPE can be implemented with the same cut-off output. However, the other way around it does not need to hold.

First order stochastic improvement. First order stochastic improvement imply that an increase in the effort of the manager M_l increases the probability to have a higher output, i.e. the cumulative distribution function is decreasing in effort. For IPE $\partial F_l(i', \theta)/\partial p_l < 0$ always holds. Under JPE $\partial F(i', \theta)/\partial p_l < 0$ only holds if the positive effect of an increase in knowledge on the output in division l outweighs the negative effect of the increase in knowledge of division l on the output of the division which supervise l. This negative effect occurs, since due to higher knowledge, M_l forwards less tasks to their superior. For the manager who is not supervised by any other manager also under JPE first order stochastic improvement always holds. In the example in the next section we show that under JPE it can be optimal to offer a contract, where a penalty occur the output of a division is higher than the target value.

JPE of the managers and direct subordinate managers. Instead of conditioning the payment of the bonus on the througput of the whole organization it can be conditioned on the

Table 6: Summary of optimal contracts for organization A, B, and C for target output $i_l \in [1, 29]$

	# Contracts			Range for target output						
	A	B	C	A Junior	B Junior	C Junior	A Middle	A Senior	B Senior	C Senior
IPE08	64	12	16	[7,10]	[7,10]	[7,10]	[8,11]	[8,11]	[9,11]	[8,11]
IPE05	64	20	16	[9,12]	[9,12]	[9,12]	[10,13]	[9,12]	[10,14]	[9,12]
JPE1	1	-	1	10	-	10	11	11	-	11
JPE2	13	-	5	$[14, 18]^a$	-	$[11, 13]^a$	$[12, 14]^a$	[10,11]	-	[10,11]

IPE08: Independent performance evaluation, intensity of queueing costs is $\beta = 0.8$
IPE05: Independent performance evaluation, intensity of queueing costs is $\beta = 0.5$
JPE1: Joint performance evaluation of a division and the division of the subordinate manager
JPE2: Joint performance evaluation of all divisions
^{a}: penalty contract

througput of the direct subordinate manager. Then first order stochastic improvement will hold

$$\frac{\partial F\left(i', \theta\right)}{\partial p_l} = -a_l \pi\left(i'_l, \theta_l\right) \prod_{m \in C_l}^{L} F_m\left(i'_m, \theta_m\right) < 0$$

Then for the same target output, the wage sensitivity to performance, B_l/A_l+B_l, is higher under JPE than IPE. This holds since $B_l^{IPE} < B_l^{JPE}$ and $A_l^{IPE} + B_l^{IPE} > A_l^{JPE} + B_l^{JPE}$ The result that the wage sensitivity to performance is higher under JPE than IPE holds for the manager who is not supervised by any other manager independent of the type of JPE.

4.4 Numerical Example Continued

We characterize the IPE and JPE contracts for the three service organizations A, B, C. In most cases the optimal contract is not unique. By changing the target output value the optimal salary and bonus payment changes. The number of optimal contracts and the range of the target output for each division is shown in table 6. We considered all possible optimal contracts with a target value of at most 29. We choose the upper bound such that the individual (joint) probability of reaching that target output is 0 with an accuracy of 4 digits after the decimal point. For each organization and each performance evaluation we discuss the payment scheme which has the lowest bonus (penalty) for all divisions. The size of the bonus payment can be interpreted as a measure for the magnitude of incentive based pay. By choosing the lowest bonus payment we compare the most conservative contract with respect to performance sensitivity. Under IPE it also coincides

Table 7: Bonus Contract for organization A,B, and C under IPE for $\beta = 0.8$

Organization	$(A,B,C)^{a,b,c}$	$A^{a,c}$	A^{b}	A^{a}, C^{a}	A^{c}, C^{b}	A^{b}	$B^{a,b}$
Rank	Junior	Middle	Middle	Senior	Senior	Senior	Senior
Salary, A_l	44.49	60.47	52.95	64.25	57.13	45.13	60.94
Bonus, B_l	22.70	27.87	34.11	28.36	34.03	44.91	31.91
Salary plus bonus, $A_l + B_l$	67.19	88.34	87.06	92.61	91.15	90.03	92.85
B_l/A_l+B_l	0.338	0.316	0.392	0.306	0.373	0.499	0.344
Target output, i'_l	10	11	10	11	10	9	11
Mode, $\lceil \theta_l \rceil - 1$	14	15	15	15	15	15	16
Expected wage, $\mathbb{E}[w_l(I)\,\vert p^*]$	59.74	79.76	79.76	83.15	83.15	83.15	84.97

[a]:B_l is minimized for all managers of the organization.
[b]:B_l/A_l+B_l is increasing in rank for all managers of the organization.
[c]:B_l/A_l+B_l is increasing between junior and senior manager.

with the payment scheme that has the highest salary and highest payment if the target output is reached.

A common measure of the importance of performance based pay is the ratio of bonus to the sum of salary and bonus, B_l/A_l+B_l, to which we will refer to as *wage sensitivity to performance* (e.g. Baiman et al. 1995). Wulf (2007) analyze the performance incentives of division managers of 250 publicly traded U.S. firms over the years 1986-1999. She reports mean measures of wage sensitivity to performance for CEOs and division managers which are increasing in the rank of the mangers. Also Baiman et al. (1995) find that the mean wage sensitivity to performance is increasing in the rank of the managers. We also discuss the payment scheme, with the lowest bonus payment possible such that the wage sensitivity to performance is increasing in the rank of the managers.

For IPE the results are summarized in Table 7. We found that the salaries and bonuses are increasing in the rank of the manager. Organization A and C have the same optimal wage contract for the junior and senior manager, respectively. The reason is that senior manager A and C only differ in their task acceptance rate which under IPE does not influence the size of the payments. In organization B the senior manager has a higher task arrival rate which results in a higher bonus payment but lower salary and the expected wage is higher. When the contracts with the lowest bonus for the senior manager are compared (Table 7, column 4 and 7) then the wage sensitivity to performance is higher in organization B than in A or C. Also under the contract with the lowest bonus payment the wage sensitivity to performance is not increasing in rank of the managers. However, by decreasing the target output, the bonus increases in order to increase the incentives

Table 8: Bonus Contract for organization A,B, and C under IPE for $\beta = 0.5$

Organization	$(A,B,C)^{a,b,c}$	A^b	A^c	A^c, C^b	A^b, C^b	B^a	B^b
Rank	Junior	Middle	Middle	Senior	Senior	Senior	Senior
Salary, A_l	45.93	56.14	62.56	61.12	51.82	68.27	61.95
Bonus, B_l	21.17	31.34	26.01	30.65	38.94	26.42	31.68
Salary plus bonus, $A_l + B_l$	67.10	87.48	88.58	91.77	90.76	94.69	93.63
Wage sensitivity, $^{B_l}/_{A_l+B_l}$	0.316	0.358	0.294	0.334	0.429	0.279	0.338
Target output, i'_l	12	12	13	12	11	14	13
Mode, $\lceil \theta_l \rceil - 1$	12	13	13	13	13	14	14
Expected wage, $\mathbb{E}[w_l(I)\,\vert p^*]$	60.40	80.70	80.70	84.18	84.18	86.80	86.80

^{a}:B_l is minimized for all managers of the organization.
^{b}:$^{B_l}/_{A_l+B_l}$ is increasing in rank for all managers of the organization.
^{c}:$^{B_l}/_{A_l+B_l}$ is increasing between junior and senior manager.

Table 9: Bonus Contract under JPE1 for $\beta = 0.8$

Organization	A,C	A	A	C
Rank	Junior	Middle	Senior	Senior
Salary, A_l	44.49	3.45	0.49	6.22
Bonus, B_l	22.70	84.89	92.12	86.39
Salary plus bonus, $A_l + B_l$	67.19	88.34	92.61	92.61
Wage sensitivity, $^{B_l}/_{A_l+B_l}$	0.338	0.961	0.995	0.933
Target output, i'_l	10	11	11	11
Mode, $\lceil \theta_l \rceil - 1$	14	15	15	15
Expected wage, $\mathbb{E}[w_l(I)\,\vert p^*]$	59.74	79.76	83.15	83.15

for a higher output.

A decrease in urgency has opposed effects on the salary and the bonus. The contracts of the managers are less sensitive to performance than under $\beta = 0.8$. We consider two different types of JPE. First, a bonus is paid to the manager if his division and the division of his subordinate manager reach a target output (JPE1). Second, a bonus is paid if all divisions reach a target output (JPE2). Under JPE1 for organization A and C the optimal contract is unique. For organization B there does not exist a contract for $i'_l \in [1, 29]$ which satisfies the limited liability condition. The results are given in table 9. The contract for the junior managers are the same as under IPE since they do not have a subordinate manager. For the middle and senior manager the composition between salary and bonus payment changes. For the middle manager the bonus payment depends on the performance of the junior manager and for the senior manager it depends on the performance of the middle manager. The salary is now decreasing in rank of the manager while the bonus payment is still increasing in rank. The difference between senior manager A and C is that the task acceptance rate

of senior manager A is higher. Additionally the subordinate manager of senior manager A (middle manager), has a lower task acceptance rate but a higher average throughput than the subordinate manager of senior manager C (junior manager). The optimal contract of senior manager A relies more on incentive based pay than the contract of senior manager C. However, the payment is the same for both managers, in case the target is reached. The difference in the bonus payments comes from the fact that a change in the cumulative distribution of the output of the division and its subordinate division resulting from higher effort is smaller for senior manager A than for senior manager C. So senior manager A has to be more incentivised.

Under JPE2 for organization B there does not exist a contract for $i_l' \in [1, 29]$ such that the limited liability condition is fulfilled. For organization A and C there only exist optimal contracts where the junior and middle manager receive a penalty if they reach the target output. This means that at the optimal effort level an increase in the effort of the manager increases the probability to have a lower joint output. Consider the division of junior manager A. An increase in the effort of the junior manager has two effects. First, an increase in the knowledge of the workers implies that the probability to solve a task is higher, and thereby increases the probability to have a higher joint output. Second, a higher level of knowledge implies less forwarded unsolved tasks to the division of the middle manager and thereby increases the probability to have a lower joint output. For both organization A and B it holds that the negative effect outweighs the positive effect. The penalty contract does not only occur because of limited liability condition but also that the managers do not have an incentive to deviate to $p_l = 0$. Under JPE2 a bonus contract would induce a manager to provide no effort at all and to rely on the effort of the other managers. The optimal penalty contract is not unique. We again chose the contract which had the lowest penalty, which is equivalent to the contract with the highest payment in case the target is reached. The target output for the junior and middle manager is higher than the mode of their individual success probability $\lceil \theta_l \rceil - 1$. For the target values optimal under IPE and JPE1 a bonus/penalty contract which implements the first best effort under limited liability does not exist under JPE2. The results for JPE2 are given in table 10. Wulf (2007) find that for division managers of lower rank, the bonus payment is typically linked to performance measures over which the manager has greater control than the overall performance of the organization. This means for the numerical example that the first type of performance evaluation is more relevant to junior and middle managers.

Table 10: Bonus Contract under JPE2 for $\beta = 0.8$

Organization	A	C	A	A	C
Rank	Junior	Junior	Middle	Senior	Senior
Salary, A_l	120.50	120.83	161.61	45.39	49.77
Bonus, B_l	-75.98	-78.41	-102.35	47.23	42.84
Salary plus bonus, $A_l + B_l$	44.52	42.42	59.26	92.61	92.61
Wage sensitivity, $^{B_l}/_{A_l+B_l}$	-1.707	-1.848	-1.727	0.509	0.46
Target output, i'_l	18	13	14	11	11
Expected wage, $\mathbb{E}\left[w_l\left(I\right)|p^*\right]$	59.74	59.74	79.76	83.15	83.15

5 Conclusion

We have studied the structure of service organizations with a special focus on control variables determined by human resource management. These are the number of workers in each division, the task acceptance rate, and the reward structure for the knowledge crew, i.e. the managers of the divisions.

We combined three features within the model. First, each manager may engage in customer contact, i.e. the task acceptance rate may be positive. We show that although the fraction of time a worker is busy is increasing in the rank of the manager, the task acceptance rate is not necessarily monotonic. The task acceptance rate of a knowledge manager depends on the task arrival rate and on the number of unsolved tasks forwarded by direct subordinate knowledge managers. In general, a knowledge manager of higher rank has a lower task acceptance rate. However, if a knowledge manager of lower rank receives comparably more unsolved problems it can occur that his task acceptance rate is lower than for the knowledge manager of higher rank. In the numerical example we discuss the case, in which it is optimal that a middle manager has a lower task acceptance rate than the senior manager. The middle manager is mainly in charge of dealing with unsolved tasks of junior managers.

Second, knowledge management by managers is explicitly taken into account. The amount of knowledge acquired by workers depends on the effort of the knowledge crew. We show that the optimal variable costs of knowledge creation per worker incurred by the knowledge manager of each division is increasing in the rank of the manager.

Third, since the effort of the knowledge crew is not easily measurable and induces costs for the knowledge manager we discuss the characteristics of a bonus contract under which it is optimal

to implement the effort level under symmetric information. We analyze two different types of reward structures for the managers. Under independent performance evaluation (IPE) only the own output influences the payment of the bonus, while under joint performance evaluation (JPE) the performance of other divisions are relevant as well. We identify conditions, which depend crucially on the hazard rate of effort, under which a bonus contract exists. The hazard rate of effort gives the change in the cumulative distribution of output resulting from higher effort in relation to the probability to receive a bonus. We find that under IPE the bonus is unambiguously positive, while under JPE instead of a bonus payment a penalty could occur. A reason is that under JPE an increase in knowledge of a worker has two effects. First, it increases the expected output of that division. Second, it decreases the output of the divisions, which supervise that division, since less unsolved tasks are forwarded. If the second effect outweighs the first effect, it is less likely that a certain overall output is achieved.

If queueing costs of the organization increase, e.g. the organization has to respond more quickly, we find the knowledge of the workers and the number of workers in each division may be substitutes. If the marginal revenue with respect to the expected number of solved tasks is elastic, it is optimal to increase the knowledge of the workers and to decrease the span of control of the knowledge manager. This implies that decisions are more decentralized, since the probability to forward a task to the superior knowledge manager decreases. An increase in urgency has opposed effects on the salary and the bonus under IPE. The contracts of the managers become more sensitive to performance.

So far the total number of divisions is taken as given. Derivation of the optimal number of divisions would allow to link the discussed control variables with the flatness of hierarchies. Nevertheless, in this model, for a given number of divisions, different organizational structures can be compared. In a numerical example we analyze a flattened firm. Initially, the knowledge crew consisted of a junior, a middle, and a senior manager. Then the middle management division is transformed into a junior management division or laid off. When middle management is transformed into junior management, we find that an task overload problem for senior managers occurs. A flattened firm may respond more quickly to changes by pushing decisions downwards. However, we find that first, the mean response time, which is the time a task spends in the organization, is higher. The reason is that under the initial knowledge crew the task acceptance rate of the senior

manager is higher and external tasks handled directly by the senior manager are not forwarded to any other manager and therefore are served faster. Second, since in the flattened organization the senior division has to solve comparably more unsolved tasks from subordinate divisions the senior manager is more involved in internal tasks. We also calculated the optimal wage contracts and find that for all managers the wage sensitivity to performance is higher under a joint performance evaluation. Also depending if the flattening occurs through lay offs or transformation result in different optimal wage contracts for the managers.

Appendix

A Hierarchy

The supervisory relationships imply the rank of the divisions. In order to assign to each division a rank we use the counting up rank system (Beckmann 1988). The rank function is given by

$$\kappa(l) = \max_{k \in C_l} \delta_{lk} \tag{17}$$

where δ_{lk} is the number of links in the path from l to k. The rank function assigns to each manager of a division a positive integer. If $\kappa(l) = r$ then M_l has rank r. If $\kappa(l) < \kappa(k)$ then M_k is in higher position than division M_l. The rank system of the organization satisfies the following properties:

1. If M_k controls M_l then $\kappa(k) > \kappa(l)$.
2. If $\left\{l \in \bar{L} : \kappa(l) = r\right\} = \emptyset$ then $\left\{l \in \bar{L} : \kappa(l) = r'\right\} = \emptyset$ for $r' > r$. So there are no gaps in the rank system.
3. $\min_{l \in \bar{L}} \{\kappa(l)\} = 0$. So the rank system is normalized at the bottom.

Let L_r be the set of managers with rank r. Then the divisions are labelled from 1 to L from left to right within each rank. The divisions with rank zero are labelled from 1 to L_1, the division with rank one are labelled from $L_1 + 1$ to L_2 and so on. The head division has label L. For an example see Figure 2(b). There the manager of division 1 and 2 have rank 0, so $L_0 = \{M_1, M_2\}$ and $\kappa(1) = \kappa(2) = 0$. For $l = 3$ it holds that $\kappa(3) = \kappa(4) = 1$.

B Properties of the routing matrix P

$P = \left(I - \overline{P}\right)^{-1}$ has the following properties:

1. P is unitriangular since $I - \overline{P}$ is unitriangular by construction.

$$P_{lk} = \begin{cases} 1 & \text{for } k = l \\ 0 & \text{for } k \notin C_l \\ \in (0,1) & \text{for } k \in C_l \end{cases} \tag{18}$$

2. P exists since $det\,(P) = det\left(I - \overline{P}\right) = 1$.

3. $a_k = \sum_m^L P_{km} t_m$ and therefore

$$\frac{\partial a_k}{\partial t_l} = P_{kl} = \begin{cases} 1 & \text{for } k = l \\ 0 & \text{for } l \notin C_k \\ \in (0,1) & \text{for } l \in C_k \end{cases} \quad \text{and} \quad \frac{\partial a_k}{\partial p_l} = \begin{cases} 0 & \text{for } k = l \\ 0 & \text{for } l \notin C_k \\ < 0 & \text{for } l \in C_k \end{cases} \tag{19}$$

C Properties of the success probability

For the network of queues considered it holds that under the equilibrium distribution, the external departure of tasks also follow independent Poisson processes (Jackson 1963). So the external departure or output follows also a Poisson process with rate $\theta_l = p_l a_l$. The probability that division l has an output of i_l to which we refer as the *success probability of division* l is given by

$$\pi\left(i_l, \theta_l\right) = e^{-\theta_l} \frac{\theta_l^{i_l}}{i_l!} \tag{20}$$

Since the external departure processes are independent the joint probability that the output is $i = (i_1, \ldots, i_L)$ is

$$\Pi\left(i, \theta\right) = \prod_{l=1}^{L} \pi\left(i_l, \theta_l\right) = e^{-\sum_{l=1}^{L} \theta_l} \frac{\theta_1^i \theta_2^i \ldots \theta_L^i}{i_1! i_2! \ldots i_L!} \tag{21}$$

The cumulative distribution of the individual success probability is

$$F_l\left(x_l, \theta_l\right) = \sum_{i_l=0}^{x_l} \pi\left(i_l, \theta_l\right)$$

The cumulative distribution of the joint success probability is

$$F\left(x, \theta\right) = \prod_{l=1}^{l} F_l\left(x_l, \theta_l\right)$$

Some remarks about the probability distributions:

Remark 1. *$\pi\left(i_l, \theta_l\right)$ is increasing for $i_l \in \{0, \ldots, \lceil \theta_l \rceil - 1\}$ and decreasing for $i_l \in \{\lceil \theta_l \rceil, \ldots\}$, where $\lceil \theta_l \rceil \in \{1, \ldots, r_l\}$. If $\lceil \theta_l \rceil = \theta_l$ then $\pi\left(i_l', \theta_l\right) = \pi\left(i_l' - 1, \theta_l\right)$.*

Proof. $\pi(i_l, \theta_l)$ is increasing iff

$$\pi(i_l, \theta_l) > \pi(i_l - 1, \theta_l) \leftrightarrow e^{-\theta_l} \frac{\theta_l^{i_l}}{i_l!} > e^{-\theta_l} \frac{\theta_l^{i_l - 1}}{(i_l - 1)!} \leftrightarrow \theta_l > i_l \tag{22}$$

If $\lceil \theta_l \rceil \neq \theta_l$ then it holds that $\theta_l > i_l$ for $i_l \leq \lceil \theta_l \rceil - 1$ and $\theta_l < i_l$ for $i_l \geq \lceil \theta_l \rceil$. If $\lceil \theta_l \rceil = \theta_l$ then $e^{-\theta_l} \frac{\theta_l^{\lceil \theta_l \rceil}}{\lceil \theta_l \rceil!} = e^{-\theta_l} \frac{\theta_l^{\lceil \theta_l \rceil - 1}}{(\lceil \theta_l \rceil - 1)!}$ or $\pi(\lceil \theta_l \rceil, \theta_l) = \pi(\lceil \theta_l \rceil - 1, \theta_l)$. Such an $\lceil \theta_l \rceil$ exists, since $0 < \theta_l = p_l a_l < p_l r_l \leq r_l$ and in the optimum the constraints $a_l < r_l$ and $p_l \in [0, 1]$ have to hold. □

Remark 2. *The first derivative of $\pi(i_k, \theta_k)$ w.r.t. p_l for $k \neq l$ is*

$$\frac{\partial \pi(i_k, \theta_k)}{\partial p_l} = \begin{cases} \frac{\partial \theta_k}{\partial p_l} \left(\pi(i_k - 1, \theta_k) - \pi(i_k, \theta_k) \right) & \text{for } i_k \in \{1, \ldots, \} \\ -\frac{\partial \theta_k}{\partial p_l} \pi(0, \theta_k) & \text{for } i_k = 0 \end{cases} \tag{23}$$

For $i_k \in \{0, \ldots, \lceil \theta_k \rceil - 1\}$ $\frac{\partial \pi(i_k, \theta_k)}{\partial p_l} > 0$ and for $i_k \in \{\lceil \theta_k \rceil, \ldots\}$ $\frac{\partial \pi(i_k, \theta_k)}{\partial p_l} < 0$. If $\lceil \theta_k \rceil = \theta_k$ then $\frac{\partial \pi(\lceil \theta_k \rceil, \theta_k)}{\partial p_l} = 0$.

The first derivative of $\pi(i_l, \theta_l)$ w.r.t. p_l is

$$\frac{\partial \pi(i_l, \theta_l)}{\partial p_l} = \begin{cases} a_l \left(\pi(i_l - 1, \theta_l) - \pi(i_l, \theta_l) \right) & \text{for } i_l \in \{1, \ldots, \} \\ -a_l \pi(0, \theta_l) & \text{for } i_l = 0 \end{cases} \tag{24}$$

For $i_l \in \{0, \ldots, \lceil \theta_l \rceil - 1\}$ $\frac{\partial \pi(i_l, \theta_l)}{\partial p_l} < 0$ and for $i_l \in \{\lceil \theta_l \rceil, \ldots\}$ $\frac{\partial \pi(i_l, \theta_l)}{\partial p_l} > 0$. If $\lceil \theta_l \rceil = \theta_l$ then $\frac{\partial \pi(\lceil \theta_l \rceil, \theta_l)}{\partial p_l} = 0$.

Proof.

$$\frac{\partial \pi(i_k, \theta_k)}{\partial p_l} = -\frac{\partial \theta_k}{\partial p_l} e^{-\theta_k} \frac{\theta_k^{i_k}}{i_k!} + e^{-\theta_k} i_k \frac{\theta_k^{i_k - 1}}{i_k!} \frac{\partial \theta_k}{\partial p_l} = \frac{\partial \theta_k}{\partial p_l} \left(\pi(i_k - 1, \theta_k) - \pi(i_k, \theta_k) \right) \text{for } i_k \in \{1, \ldots\} \tag{25}$$

and

$$\frac{\partial \pi(0, \theta_k)}{\partial p_l} = -\frac{\partial \theta_k}{\partial p_l} e^{-\theta_k} = -\frac{\partial \theta_k}{\partial p_l} \pi(0, \theta_k) \tag{26}$$

Note that

$$\frac{\partial \theta_k}{\partial p_l} = \begin{cases} a_l & \text{for } k = l \\ 0 & \text{for } l \notin C_k \\ p_k \frac{\partial a_k}{\partial p_l} & \text{for } l \in C_k \end{cases}$$

The signs of the derivative follow from remark 1 and from $\frac{\partial a_k}{\partial p_l} < 0$ for $l \in C_k$. □

Remark 3. *The first derivative of the cumulative distribution $F_k(x_k, \theta_l)$ w.r.t. p_l is*

$$\frac{\partial F_k(x_k, \theta_l)}{\partial p_l} = \begin{cases} -a_l \pi(x_l, \theta_l) < 0 & \text{for } k = l \\ 0 & \text{for } l \notin C_k \\ -\frac{\partial \theta_k}{\partial p_l} \pi(x_k, \theta_k) > 0 & \text{for } l \in C_k \end{cases} \tag{27}$$

Proof. If $k = l$, then for $x_l = 0$ it holds by remark 2. For $x_l > 0$

$$\frac{\partial F_l(x_l, \theta_l)}{\partial p_l} = \sum_{i_l=0}^{x_l} \frac{\partial \pi(i_l, \theta_l)}{\partial p_l} = -a_l \pi(0, \theta_l) + \sum_{i_l=1}^{x_l} a_l (\pi(i_l - 1, \theta_l) - \pi(i_l, \theta_l)) = -a_l \pi(x_l, \theta_l) < 0 \tag{28}$$

If $l \in C_k$, then for $x_k = 0$ it holds by remark 2. For $x_k > 0$

$$\sum_{i_k=0}^{x_k} \frac{\partial \pi(i_k, \theta_k)}{\partial p_l} = -\frac{\partial \theta_k}{\partial p_l} \pi(0, \theta_k) + \sum_{i_k=1}^{x_k} \frac{\partial \theta_k}{\partial p_l} (\pi(i_k - 1, \theta_k) - \pi(i_k, \theta_k)) = -\frac{\partial \theta_k}{\partial p_l} \pi(x_k, \theta_k) > 0$$

If $l \notin C_k$, then $\frac{\partial F_k(x_k, \theta_k)}{\partial p_l} = 0$ since $\frac{\partial \theta_k}{\partial p_l} = 0$. □

Remark 4. *Let $x_l < \lceil \theta_l \rceil$. Then*

$$\frac{\partial^2 F_l(x_l, \theta_l)}{\partial p_l^2} > 0$$

Proof.

$$\frac{\partial^2 F_l(x_l, \theta_l)}{\partial p_l^2} = \frac{\partial (-a_l \pi(x_l, \theta_l))}{\partial p_l} = -a_l \frac{\partial \pi(x_l, \theta_l)}{\partial p_l} > 0 \tag{29}$$

where the sign follows by remark 2. □

Remark 5. *The first derivative of the joint cumulative distribution $F(x,\theta)$ w.r.t. p_l is*

$$\frac{\partial F(x,\theta)}{\partial p_l} = -\sum_{k:l\in C_k} F_{-k}(x,\theta)\frac{\partial \theta_k}{\partial p_l}\pi(x_k,\theta_k) - F_{-l}(x,\theta)\, a_l \pi(x_l,\theta_l) \tag{30}$$

where $F_{-k}(x,\theta) = \prod_{l\neq k}^{L} F_l(x_i,\theta_l)$.

Proof. Since $F(x,\theta) = \prod_{l=1}^{L} F_l(x_l,\theta_l)$ it follows that

$$\begin{aligned}\frac{\partial F(x,\theta)}{\partial p_l} &= \sum_{k=1}^{L} F_{-k}(x,\theta)\frac{\partial F_k(x_k,\theta_l)}{\partial p_l} \\ &= -\sum_{k:l\in C_k} F_{-k}(x,\theta)\frac{\partial \theta_k}{\partial p_l}\pi(x_k,\theta_k) - F_{-l}(x,\theta)\, a_l\pi(x_l,\theta_l)\end{aligned}$$

where the first part is positive and the second part is negative. □

Remark 6. *For the second derivative of the joint cumulative distribution $F(x,\theta)$ w.r.t. p_l it holds that*

1. *If $\partial F(x,\theta)/\partial p_l < 0$ and $x_l \leq \lceil \theta_l \rceil - 1$ then $\partial^2 F(x,\theta)/\partial p_l^2 > 0$.*
2. *If $\partial F(x,\theta)/\partial p_l > 0$ and $x_l \geq \lceil \theta_l \rceil - 1$ then $\partial^2 F(x,\theta)/\partial p_l^2 < 0$.*

Proof.

$$\begin{aligned}\frac{\partial^2 F(x,\theta)}{\partial p_l^2} =& -\sum_{k:l\in C_k}\frac{\partial F_{-k}(x,\theta)}{\partial p_l}\frac{\partial \theta_k}{\partial p_l}\pi(x_k,\theta_k) - \sum_{k:l\in C_k} F_{-k}(x,\theta)\frac{\partial \theta_k}{\partial p_l}\frac{\partial \pi(x_k,\theta_k)}{\partial p_l} \\ &- \frac{\partial F_{-l}(x,\theta)}{\partial p_l} a_l\pi(x_l,\theta_l) - F_{-l}(x,\theta)\, a_l \frac{\partial \pi(x_l,\theta_l)}{\partial p_l} \\ =& \sum_{k:l\in C_k}\frac{\partial \theta_k}{\partial p_l}\pi(x_k,\theta_k)\left(\sum_{s:l\in C_s, s\neq k} F_{-s,k}(x,\theta)\frac{\partial \theta_s}{\partial p_l}\pi(x_s,\theta_s) + F_{-l,k}(x,\theta)\, a_l\pi(x_l,\theta_l)\right) \\ &- \sum_{k:l\in C_k} F_{-k}(x,\theta)\left(\frac{\partial \theta_k}{\partial p_l}\right)^2 (\pi(x_k-1,\theta_k,) - \pi(x_k,\theta_k)) \\ &+ a_l\pi(x_l,\theta_l)\left(\sum_{k:l\in C_k,} F_{-k,l}(x,\theta)\frac{\partial \theta_k}{\partial p_l}\pi(x_k,\theta_k)\right) - F_{-l}(x,\theta)\, a_l^2(\pi(x_l-1,\theta_l) - \pi(x_l,\theta_l)) \\ =& \sum_{k:l\in C_k}\frac{\partial \theta_k}{\partial p_l}\pi(x_k,\theta_k)\left(2a_l\pi(\theta_l,x_l)\, F_{-k,l}(x,\theta) + \sum_{s:l\in C_s, s\neq k} F_{-s,k}(x,\theta)\frac{\partial \theta_s}{\partial p_l}\pi(x_s,\theta_s)\right) \\ &- \sum_{k:l\in C_k} F_{-k}(x,\theta)\left(\frac{\partial \theta_k}{\partial p_l}\right)^2 (\pi(x_k-1,\theta_k) - \pi(x_k,\theta_k)) - F_{-l}(x,\theta)\, a_l^2(\pi(x_l-1,\theta_l) - \pi(x_l,\theta_l))\end{aligned}$$

1.

2. If multiplying through one can see that all terms except the second term in the first line are negative. If $\frac{\partial F(x,\theta)}{\partial p_l} > 0$, then it holds that for any $k : l \in C_k, k \neq s$

$$a_l \pi\left(x_l, \theta_l\right) F_{-l,k}\left(x, \theta\right) < -\sum_{s: l \in C_s} F_{-s,k}\left(x, \theta\right) \frac{\partial \theta_s}{\partial p_l} \pi\left(x_s, \theta_s\right)$$

So

$$2a_l \pi\left(x_l, \theta_l\right) F_{-k,l}\left(x, \theta\right) + \sum_{s: l \in C_s, s \neq k} F_{-s,k}\left(x, \theta\right) \frac{\partial \theta_s}{\partial p_l} \pi\left(x_s, \theta_s\right) < -\sum_{s: l \in C_s, s \neq k} F_{-s,k}\left(x, \theta\right) \frac{\partial \theta_s}{\partial p_l} \pi\left(x_s, \theta_s\right)$$

and therefore for the first line it holds that

$$\sum_{k: l \in C_k} \frac{\partial \theta_k}{\partial p_l} \pi\left(x_k, \theta_k\right) \left(2a_l \pi\left(x_l, \theta_l\right) F_{-k,l}\left(x, \theta\right) + \sum_{s: l \in C_s, s \neq k} F_{-s,k}\left(x, \theta\right) \frac{\partial \theta_s}{\partial p_l} \pi\left(x_s, \theta_s\right) \right) <$$
$$- \sum_{k: l \in C_k} \frac{\partial \theta_k}{\partial p_l} \pi\left(x_k, \theta_k\right) \left(\sum_{s: l \in C_s, s \neq k} F_{-s,k}\left(x, \theta\right) \frac{\partial \theta_s}{\partial p_l} \pi\left(x_s, \theta_s\right) \right) < 0$$

Therefore $\frac{\partial^2 F(x,\theta)}{\partial p_l^2} < 0$ if $\frac{\partial F(x,\theta)}{\partial p_l} > 0$ and $x_l \geq \lceil \theta_l \rceil - 1$.

□

D Key performance indicators of the organization

For the network of queues considered it holds that under the equilibrium distribution, the external departure of tasks also follow independent Poisson processes (Jackson 1963). Therefore the throughput of division l is

$$\mathbb{E}\left[I_l | t\right] = \sum_{i_l} \pi\left(i_l, \theta_l\right) i_l = \theta_l \tag{31}$$

and the total expected throughput is

$$\mathbb{E}\left[I | t\right] = \mathbb{E}\left[\sum_{l=1}^{L} I_l | t\right] = \sum_{l=1}^{L} \sum_{i_l} \pi\left(i_l, \theta_l\right) i_l = \sum_{l=1}^{L} \theta_l \tag{32}$$

Let $n_l = (n_{l1}, \ldots, n_{lr_l})$ be the number of tasks pending and in service in division l. From the Jackson Theorem (Jackson 1963) it follows that the probability that the overall system state is $(n_1, \ldots, n_l)$ has a product form expression. So the performance of each division can be analysed independently as a $M/M/1$ queue for each worker. The mean number of tasks in the system is the sum of the tasks at every division, pending and in service

$$N = \sum_{l=1}^{L} N_l = \sum_{l=1}^{L} \sum_{k=1}^{r_l} n_{lk} \mathbb{P}\left(n_{lk} \text{ tasks at worker } k \text{ in division } l\right) = \sum_{l=1}^{L} r_l \frac{\rho_l}{1-\rho_l} = \sum_{l=1}^{L} \frac{a_l}{1-\rho_l}$$

where ρ_l is the traffic intensity and measures the mean number of tasks in service at a worker in division l. $\rho_l/1-\rho_l$ is the mean number of tasks at a worker in division l, queued and in service. The mean number of tasks pending in the system is the total number of tasks minus the tasks in service

$$Q = \sum_{l=1}^{L} Q_l = \sum_{l=1}^{L} r_l \left(\frac{\rho_l}{1-\rho_l} - \rho_l \right) = \sum_{l=1}^{L} \frac{a_l^2}{r_l - a_l} \tag{33}$$

Note that $Q_l/N_l = \rho_l$. Since the service rate is equal to one, the traffic intensity is equal to the fraction of queued tasks on total tasks in the division. The mean response time is the time a task spends in the organization and can be derived from Little's formula. It states that the mean number of tasks in the system have to be equal to the sum of all external arrival rates (the task acceptance rates) times the mean response time.

$$W = \frac{\sum_{l=1}^{L} \frac{a_l}{1-\rho_l}}{\sum_{l=1}^{L} t_l}$$

E Proofs for the Organizational Structure

E.1 First order conditions

The constrained optimization problem is transformed into an unconstrained problem by plugging condition BE into the objective function.

$$\max_{t,p,r} \sum_{l=1}^{L} \left(H_l\left(\theta_l\right) - \beta \frac{a_l^2}{r_l - a_l} - G_l\left(p_l\right) - c_l^M r_l - c_l^W r_l \right)$$

where $a_l = \sum_{m=1}^{l} P_{lm} t_m$. The first order conditions for t_l is

$$\sum_k H_\theta(\theta_l) \frac{\partial a_k}{\partial t_l} p_k - \beta \frac{2a_k \frac{\partial a_k}{\partial t_l}(r_k - a_k) - (a_k)^2 \left(-\frac{\partial a_k}{\partial t_l}\right)}{(r_k - a_k)^2} = 0$$

$$\sum_{k: l \in C_k} P_{kl} \left(H_\theta(\theta_l) p_k - \beta \frac{2a_k r_k - (a_k)^2}{(r_k - a_k)^2} \right) = 0$$

Since the head division of the organization L is not controlled by any other division, i.e. $L \notin C_k$ for any k, it follows that $P_{kL} = 0$ for $k \neq L$

$$H_\theta(\theta_L^*) p_L^* - \beta \frac{2a_L^* r_L^* - (a_L^*)^2}{(r_L^* - a_L^*)^2} = 0 \tag{34}$$

Since by assumption each division forwards only to one other division, solving recursively for $l \in \{L_{L-1}, \ldots, L_0\}$ gives

$$H_\theta(\theta_l^*) p_l^* = \beta \frac{2a_l^* r_l^* - (a_l^*)^2}{(r_l^* - a_l^*)^2} \text{ for } l = 1, \ldots, L-1 \tag{35}$$

The first order conditions for p_l is

$$\sum_{k=1}^{L} \left(H_\theta(\theta_l) \frac{d\theta_k}{dp_l} - \beta \frac{2a_k \frac{\partial a_k}{\partial p_l}(r_k - a_k) - (a_k)^2 \left(-\frac{\partial a_k}{\partial p_l}\right)}{(r_k - a_k)^2} \right) - g_p(p_l) = 0$$

$$a_l + \sum_{k=1}^{L} \frac{\partial a_k}{\partial p_l} \left(H_\theta(\theta_l) p_k - \beta \frac{2a_k r_k - (a_k)^2}{(r_k - a_k)^2} \right) - g_p(p_l) = 0$$

$$H_\theta(\theta_l^*) a_l^* - g_p(p_l^*) = 0$$

The first order condition for r_l is

$$\frac{\beta (a_l)^2}{(r_l - a_l)^2} - c_l^W - c_l^M = 0$$

Solving for r_l gives

$$(r_l - a_l)^2 = \frac{\beta}{c_l^W + c_l^M} (a_l)^2$$
$$r_l^* = a_l^* \left(1 + \sqrt{\frac{\beta}{c_l^W + c_l^M}}\right) > a_l^*$$

The second order condition is fulfilled if the Hessian matrix is negative definite. Let

$$F(t, p, r) = \sum_{l=1}^{L} \left(H_l(\theta_l) - \beta \frac{a_l^2}{r_l - a_l} - G_l(p_l) - c_l^M r_l - c_l^W r_l \right)$$

Then the Hessian matrix is given by

$$H = \begin{bmatrix} H_1 & H_2 \\ H_2^T & H_3 \end{bmatrix}$$

where

$$H_1 = \begin{bmatrix} \frac{\partial^2 F}{\partial t_1 \partial t_1} & \cdots & \frac{\partial^2 F}{\partial t_1 \partial t_L} & \frac{\partial^2 F}{\partial t_1 \partial p_1} & \cdots & \frac{\partial^2 F}{\partial t_1 \partial p_L} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 F}{\partial t_L \partial t_1} & \cdots & \frac{\partial^2 F}{\partial t_L \partial t_L} & \frac{\partial^2 F}{\partial t_L \partial p_1} & \cdots & \frac{\partial^2 F}{\partial t_L \partial p_L} \\ \frac{\partial^2 F}{\partial p_1 \partial t_1} & \cdots & \frac{\partial^2 F}{\partial p_1 \partial t_L} & \frac{\partial^2 F}{\partial p_1 \partial p_1} & \cdots & \frac{\partial^2 F}{\partial p_1 \partial p_L} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 F}{\partial p_L \partial t_1} & \cdots & \frac{\partial^2 F}{\partial p_L \partial t_L} & \frac{\partial^2 F}{\partial p_L \partial p_1} & \cdots & \frac{\partial^2 F}{\partial p_L \partial p_L} \end{bmatrix}, H_2 = \begin{bmatrix} \frac{\partial^2 F}{\partial t_1 \partial r_1} & \cdots & \frac{\partial^2 F}{\partial t_1 \partial r_L} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 F}{\partial t_L \partial r_1} & \cdots & \frac{\partial^2 F}{\partial t_L \partial r_L} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 F}{\partial p_1 \partial r_1} & \cdots & \frac{\partial^2 F}{\partial p_1 \partial r_L} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 F}{\partial p_L \partial r_1} & \cdots & \frac{\partial^2 F}{\partial p_L \partial r_L} \end{bmatrix}, H_3 = \begin{bmatrix} \frac{\partial^2 F}{\partial r_1 \partial r_1} & \cdots & \frac{\partial^2 F}{\partial r_1 \partial r_L} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 F}{\partial r_L \partial r_1} & \cdots & \frac{\partial^2 F}{\partial r_L \partial r_L} \end{bmatrix}$$

In order to check if the Hessian is negative definite, we use the following result: For any symmetric matrix, M, of the form

$$M = \begin{bmatrix} M_1 & M_2 \\ M_2^T & M_3 \end{bmatrix}$$

if M_3 is invertible then

1. M is positive definite iff M_3 is positive definite and $M_1 - M_2 M_3^{-1} M_2^T$ is positive definite.

2. M is positive definite iff M_1 is positive definite and $M_3 - M_2^T M_1^{-1} M_2$ is positive definite.

In order to apply this result, let $M = -H$, $M_1 = -H_1$, $M_2 = -H_2$ and $M_3 = -H_3$. $-H_3$ is a diagonal matrix and invertible since ${}^{2\beta a_l^2}/_{(r_l - a_l)^3} > 0$. After some matrix multiplication we find that

$$M_1 - M_2 M_3^{-1} M_2^T = \begin{bmatrix} M_1' & M_2' \\ M_2'^T & M_3' \end{bmatrix} = \begin{bmatrix} -P^T E P & -P^T \left(EA + C\right) \\ -\left(A^T E + C\right) P & -A^T E A - A^T C - CA - D \end{bmatrix}$$

where C, D, E are a $L \times L$ matrix with $H_{\theta\theta} a_l p_l + H_\theta$, $H_{\theta\theta} a_l^2 - g_{pp}$, and $H_{\theta\theta} p_l^2$ on the diagonal, respectively and

$$A = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \frac{\partial a_2}{\partial p_1} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\partial a_L}{\partial p_1} & \dots & \frac{\partial a_L}{\partial p_{L-1}} & 0 \end{bmatrix}, P = \begin{bmatrix} P_{11} & 0 & \dots & 0 \\ P_{21} & P_{22} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ P_{L1} & \dots & P_{LL-1} & P_{LL} \end{bmatrix}$$

In order to determine if $M_1 - M_2 M_3^{-1} M_2^T$ is positive definite we apply the second result. $-P^T E P$ is invertible and positive definite if $-E$ is positive definite. This holds since $-H_{\theta\theta} p_l^2 > 0$. $M_3' - M_2'^T M_1'^{-1} M_2'$ is positive definite if $-D + CE^{-1}C$ is positive definite, since

$$\begin{aligned} -A^T E A - A^T C - CA - D - \left(-P^T \left(EA + C\right)\right)^T \left(-P^T E P\right)^{-1} \left(-P^T \left(EA + C\right)\right) = \\ -A^T E A - A^T C - CA - D + \left(A^T E + C\right) P P^{-1} E^{-1} \left(P^T\right)^{-1} P^T \left(EA + C\right) = \\ -A^T E A - A^T C - CA - D + A^T E A + CA + A^T C + CE^{-1}C = -D + CE^{-1}C \end{aligned}$$

$$-D_{ll} + C_{ll} E_{ll}^{-1} C_{ll} = -H_{\theta\theta} a_l^2 + g_{pp} + \frac{\left(H_{\theta\theta} a_l p_l + H_\theta\right)^2}{H_{\theta\theta} p_l^2} > 0 \text{ if } 2H_{\theta\theta} a_l p_l + H_\theta < 0$$

$M_1 - M_2 M_3^{-1} M_2^T$ is positive definite if $2H_{\theta\theta} a_l p_l + H_\theta < 0$ and therefore the Hessian is negative definite if $2H_{\theta\theta} a_l p_l + H_\theta < 0$.

E.2 Comparative Static

The parameter β measures the importance of the number of tasks pending in the organization and $c_l = c_l^W + c_l^M$ are the total costs for workers. In this section the effect of β and c_l^W on p_l^*, r_l^*, t_l^* and ρ_l^* are derived. From the first order conditions we can derive two equations in p_l^* and r_l^* and apply the implicit function theorem.

$$F^1(p_l, r_l, \beta, c_l) = H_\theta p_l - \left(c_l + 2\sqrt{\beta c_l}\right) = 0$$
$$F^2(p_l, r_l, \beta, c_l) = H_\theta a_l - g_p = 0$$

with $a_l = \frac{r_l\sqrt{c_l}}{\sqrt{c_l}+\sqrt{\beta}}$ and as a short hand notation $H_\theta(\theta_l) = H_\theta$ and $g_p(p_l) = g_p$ The relevant derivates in order to apply the implicit function theorem are

$$\begin{aligned}
\frac{\partial F^1}{\partial p_l} &= H_{\theta\theta} a_l p_l + H_\theta & \frac{\partial F^2}{\partial p_l} &= H_{\theta\theta}(a_l)^2 - g_{pp} \\
\frac{\partial F^1}{\partial r_l} &= H_{\theta\theta}(p_l)^2 \frac{\partial a_l}{\partial r_l} & \frac{\partial F^2}{\partial r_l} &= (H_{\theta\theta} a_l p_l + H_\theta)\frac{\partial a_l}{\partial r_l} \\
\frac{\partial F^1}{\partial \beta} &= H_{\theta\theta}(p_l)^2 \frac{\partial a_l}{\partial \beta} - \sqrt{\frac{c_l}{\beta}} & \frac{\partial F^2}{\partial \beta} &= (H_{\theta\theta} a_l p_l + H_\theta)\frac{\partial a_l}{\partial \beta} \\
\frac{\partial F^1}{\partial c_l} &= H_{\theta\theta}(p_l)^2 \frac{\partial a_l}{\partial c_l} - 1 - \sqrt{\frac{\beta}{c_l}} & \frac{\partial F^2}{\partial c_l} &= (H_{\theta\theta} a_l p_l + H_\theta)\frac{\partial a_l}{\partial c_l}
\end{aligned}$$

where

$$\frac{\partial a_l}{\partial r_l} = \frac{\sqrt{c_l}}{\sqrt{c_l}+\sqrt{\beta}} > 0, \quad \frac{\partial a_l}{\partial \beta} = -\frac{r_l\sqrt{c_l}}{2\sqrt{\beta}\left(\sqrt{c_l}+\sqrt{\beta}\right)^2} < 0, \quad \frac{\partial a_l}{\partial c_l} = \frac{r_l\sqrt{\beta}}{2\sqrt{c_l}\left(\sqrt{c_l}+\sqrt{\beta}\right)^2} > 0$$

Then

$$J = \det\left(\begin{bmatrix} \frac{\partial F^1}{\partial p_l} & \frac{\partial F^1}{\partial r_l} \\ \frac{\partial F^2}{\partial p_l} & \frac{\partial F^2}{\partial r_l} \end{bmatrix}\right) = \left(H_\theta\left(H_\theta + 2H_{\theta\theta}\theta_l\right) + H_{\theta\theta} p_l^2 g_{pp}\right)\frac{\partial a_l}{\partial r_l} < 0 \tag{36}$$

The second order conditions of the optimization problem is fulfilled if $-{}^{H_{\theta\theta}\theta_l}/_{H_\theta} > 0.5$. Therefore $H_\theta + 2H_{\theta\theta}\theta_l < 0$ and so $J < 0$. If $-{}^{H_{\theta\theta}\theta_l}/_{H_\theta} \geq 1$ then

$$J_{p\beta} = \det\left(\begin{bmatrix} \frac{\partial F^1}{\partial \beta} & \frac{\partial F^1}{\partial r_l} \\ \frac{\partial F^2}{\partial \beta} & \frac{\partial F^2}{\partial r_l} \end{bmatrix}\right) = -\left(H_\theta + H_{\theta\theta}\theta_l\right)\sqrt{\frac{c_l}{\beta}}\frac{\partial a_l}{\partial r_l} \geq 0$$

$$J_{pc} = \det\left(\begin{bmatrix} \frac{\partial F^1}{\partial c_l} & \frac{\partial F^1}{\partial r_l} \\ \frac{\partial F^2}{\partial c_l} & \frac{\partial F^2}{\partial r_l} \end{bmatrix}\right) = -\left(H_\theta + H_{\theta\theta}\theta_l\right)\left(1 + \sqrt{\frac{\beta}{c_l}}\right)\frac{\partial a_l}{\partial r_l} \geq 0$$

$$J_{rc} = \det\left(\begin{bmatrix} \frac{\partial F^1}{\partial p_l} & \frac{\partial F^1}{\partial c_l} \\ \frac{\partial F^2}{\partial p_l} & \frac{\partial F^2}{\partial c_l} \end{bmatrix}\right) = \left(H_{\theta\theta}p_l^2 g_{pp} + H_\theta\left(H_\theta + 2H_{\theta\theta}\theta_l\right)\right)\frac{\partial a_l}{\partial c_l} + \left(1 + \sqrt{\frac{\beta}{c_l}}\right)\left(H_{\theta\theta}a_l^2 - g_p\right) \leq 0$$

$$J_{r\beta} = \det\left(\begin{bmatrix} \frac{\partial F^1}{\partial p_l} & \frac{\partial F^1}{\partial \beta} \\ \frac{\partial F^2}{\partial p_l} & \frac{\partial F^2}{\partial \beta} \end{bmatrix}\right) = \left(H_{\theta\theta}p_l^2 g_{pp} + H_\theta\left(H_\theta + 2H_{\theta\theta}\theta_l\right)\right)\frac{\partial a_l}{\partial \beta} + \sqrt{\frac{c_l}{\beta}}\left(H_{\theta\theta}a_l^2 - g_p\right) \leq 0$$

where $J_{r\beta} \leq 0$ follows from using

$$\frac{\partial a_l}{\partial \beta} = -\frac{a_l}{2\sqrt{\beta}\left(\sqrt{c_l} + \sqrt{\beta}\right)}, \quad -p_l g_{pp} - 2g_p < {}^{H_\theta^2}/_{H_{\theta\theta}p_l}$$

where the inequality follows from equation 36. So the effect of β and c_l on p_l^* and r_l^* are

$$\frac{\partial p_l^*}{\partial \beta} \geq 0, \frac{\partial p_l^*}{\partial c_l} \geq 0, \frac{\partial r_l^*}{\partial \beta} \leq 0, \frac{\partial r_l^*}{\partial c_l} \leq 0$$

F Proofs for the Incentive Structure

F.1 Proof of Proposition 1 [IPE]

Proof. We show that under the specified contract for $p_l = p_l^*$ the condition LL is satisfied, condition IR is binding and IC is fulfilled and that p_l^* is the maximizer on $[0, 1]$. Then the contract is also optimal for HRM since he can induce the first-best effort level p_l^* by paying each manager M_l his reservation utility.

Condition (IR) is binding for $A_l > 0$: Suppose it is not binding at $p_l = p_l^*$

$$U\left(p_l\right) = A_l + \left(1 - F_l\left(i_l', \theta_l^*\right)\right)B_l - G_l\left(p_l^*\right) - c_l^M r_l^* > 0$$

Then HRM can lower the salary $A_l > 0$ by some $\epsilon > 0$ small enough such that (IR) still holds. Since B_l stays the same also the effort level will stay at p_l^*. The wage costs of the principal decreases while the revenue and the other costs stay the same. A_l could not have been optimal and (IR) has to be binding and can be used to determine the salary A_l

$$A_l = G_l\left(p_l^*\right) + c_l^M r_l^* - \left(1 - F_l\left(i_l', \theta_l^*\right)\right) B_l \tag{37}$$

Condition (IC) is replaced by the first order condition and can be used to determine the bonus B_l

$$\frac{\partial U\left(p_l\right)}{\partial p_l} = -\frac{\partial F_l\left(i_l', \theta_l\right)}{\partial p_l} B_l - g_p\left(p_l\right) = 0 \rightarrow B_l = -\frac{g_p\left(p_l^*\right)}{\frac{\partial F_l\left(i_l', \theta_l^*\right)}{\partial p_l}} \tag{38}$$

Then $p_l = p_l^*$ is a solution to the first order condition. Since $\frac{\partial F_l\left(i_l', \theta_l\right)}{\partial p_l} = -a_l \pi_l\left(i_l', \theta_l\right) < 0$ also $B_l > 0$.

Condition (LL) is fulfilled if

$$A_l = G_l\left(p_l^*\right) + c_l^M r_l^* - \left(1 - F_l\left(i_l', \theta_l^*\right)\right) B_l > 0 \leftrightarrow -\frac{\frac{\partial F_l\left(i_l', \theta_l^*\right)}{\partial p_l}}{\left(1 - F_l\left(i_l', \theta_l^*\right)\right)} > \frac{g_p\left(p_l^*\right)}{G_l\left(p_l^*\right) + c_l^M r_l^*}$$

which holds by the initial assumption.

It has to be verified if $p_l = p_l^*$ is maximizer on $[0, 1]$. The second order condition under the optimal contract is

$$-\frac{\partial^2 F_l\left(i_l', \theta_l\right)}{\partial p_l^2} B_l - g_{pp}\left(p_l^*\right) = -\left(a_l^*\right)^2 \left(\pi\left(i_l' - 1, \theta_l^*\right) - \pi\left(i_l', \theta_l^*\right)\right) B_l - g_{pp}\left(p_l^*\right) < 0$$

which follows from $\frac{\partial^2 F_l\left(i_l', \theta_l\right)}{\partial p_l^2} > 0$ for $i_l' < \theta_l^*$ (see Appendix C, remark 4). So $p_l^* > \frac{i_l'}{a_l^*}$ is a local maximizer. However, there can be more than one critical point to the first order condition. For $p_l \in \left(p_l^*, 1\right]$, since $p_l > p_l^*$ it holds that $i_l' < \theta_l$, $g_p\left(p_l\right) > g_p\left(p_l^*\right)$ and $\pi\left(\theta_l, i_l'\right) < \pi\left(\theta_l^*, i_l'\right)$. Therefore $\partial U(p_l)/\partial p_l < 0$ and no critical point can occur in this interval. For $p_l \in \left[i_l'/a_l^*, p_l^*\right)$ it holds that $g_p\left(p_l\right) < g_p\left(p_l^*\right)$ and $\pi\left(\theta_l, i_l'\right) > \pi\left(\theta_l^*, i_l'\right)$. Therefore $\partial U(p_l)/\partial p_l > 0$ and no critical point can occur in

this interval. In order for the manager not to deviate to $p_l = 0$ the utility at p_l^* has to be higher

$$\lim_{p_l \to 0} U(p_l) = A_l - c_l^M r_l^* = G_l(p_l^*) - \left(1 - F_l\left(i_l', \theta_l^*\right)\right) B_l < 0 = U(p_l^*)$$

which holds by assumption. Therefore it holds that $p_l = p_l^*$ is maximizer on $[0, 1]$. □

F.2 Proof of Proposition 2 [JPE]

Proof. We show that under the specified contract for $p_l = p_l^*$ condition LL is satisfied, condition IR is binding and IC is fulfilled and p_l^* is the maximizer on $[0, 1]$. Then the contract is also optimal for HRM since he can induce the first-best effort level p_l^* by paying each mamager M_l his reservation utility.

By the same line of reasoning as in the previous proof (IR) has to be binding and can be used to determine the salary A_l:

$$A_l = G_l(p_l^*) + c_l^M r_l^* - \left(1 - F\left(i', \theta^*\right)\right) B_l \tag{39}$$

Condition (IC) is replaced by the first order condition and can be used to determine the bonus B_l

$$\frac{\partial U(p_l)}{\partial p_l} = -\frac{\partial F(i', \theta)}{\partial p_l} B_l - g_p(p_l) = 0 \to B_l = -\frac{g_p(p_l^*)}{\frac{\partial F(i', \theta^*)}{\partial p_l}} \tag{40}$$

Then $p_l = p_l^*$ is a solution to the first order condition. For $B_l > 0$ it has to hold that

$$\frac{\partial F(i', \theta^*)}{\partial p_l} = -\sum_{k: l \in C_k} F_{-k}\left(i', \theta^*\right) \frac{\partial \theta_k}{\partial p_l} \pi\left(i_k', \theta_k^*\right) - F_{-l}\left(i', \theta^*\right) a_l \pi\left(i_l', \theta_l^*\right) < 0$$

Condition (LL) is fulfilled if

$$A_l = G_l(p_l^*) + c_l^M r_l^* + \left(1 - F\left(i', \theta_l^*\right)\right) B_l > 0 \leftrightarrow -\frac{\frac{\partial F(i', \theta^*)}{\partial p_l}}{(1 - F(i', \theta^*))} > \frac{g_p(p_l^*)}{G_l\left(p_l^*\right) + c_l^M r_l^*}$$

which holds by the initial assumption.

It has to be verified if $p_l = p_l^*$ is maximizer on $[0, 1]$. The second order condition under the

optimal contract is

$$-\frac{\partial^2 F_l\left(i_l', \theta_l^*\right)}{\partial p_l^2} B_l - g_{pp}\left(p_l^*\right) < 0$$

which follows from $\frac{\partial^2 F_l\left(i_l', \theta_l^*\right)}{\partial p_l^2} > 0$ for $i_l' < \theta_l^*$ (see Appendix C, remark 4).

In order for the manager not to deviate to $p_l = 0$ the utility at p_l^* has to be higher

$$\lim_{p_l \to 0} U\left(p_l\right) = A_l + \lim_{p_l \to 0} \left(1 - F\left(i', \theta^*\right)\right) B_l - c_l^M r_l^* < 0 = U\left(p_l^*\right)$$

which holds by assumption. □

References

Armstrong, Michael. 2003. *A handbook of human resource management practice*. Kogan Page Limited.

Baiman, Stanley, David F. Larcker, Madhav V. Rajan. 1995. Organizational design for business units. *Journal of Accounting Research* **33**(2) pp. 205–229.

Beckmann, Martin J. 1988. *Tinbergen Lectures on Organization Theory*. Texts and monographs in economics and mathematical systems, Springer.

Beggs, Alan W. 2001. Queues and hierarchies. *The Review of Economic Studies* **68**(2) 297–322.

Bowen, David E., Cheri Ostroff. 2004. Understanding HRM - firm performance linkages: The role of the strength of the hrm system. *Academy of Management Review* **29**(2) 203 – 221.

Boxall, Peter, John Purcell. 2003. *Strategy and human resource management*. Palgrave Macmillan.

Calvo, Guillermo A, Stanislaw Wellisz. 1978. Supervision, loss of control, and the optimum size of the firm. *Journal of Political Economy* **86**(5) 943–52.

Cho, Myeonghwan. 2010. Efficient structure of organization with heterogeneous workers. *Journal of Mathematical Economics* **46**(6) 1125–1139.

Colombo, Massimo G., Luca Grilli. 2013. The creation of a middle-management level by entrepreneurial ventures: Testing economic theories of organizational design. *Journal of Economics & Management Strategy* **22**(2) 390–422.

Eisenhardt, Kathleen M. 1985. Control: Organizational and economic approaches. *Management science* **31**(2) 134–149.

Gao, Fei, Meng Li, Steve Clarke. 2008. Knowledge, management, and knowledge management in business operations. *Journal of Knowledge Management* **12**(2) 3–17.

Garicano, Luis. 2000. Hierarchies and the organization of knowledge in production. *Journal of Political Economy* **108**(5) 874–904.

Goleman, Daniel. 2000. Leadership that gets results. *Harvard business review* **78**(2) 78.

Jackson, James R. 1963. Jobshop-like queueing systems. *Management Science* **10**(1) 131–142.

Kim, Son Ku. 1997. Limited liability and bonus contracts. *Journal of Economics & Management Strategy* **6**(4) 899–913.

Koys, Daniel J. 2001. The Effects of Employee Satisfaction, Organizational Citizenship Behavior, and Turnover on Organizational Effectiveness: a Unit-Level, Longitudinal Study. *Personnel Psychology* **54**(1) 101–114.

Larsson, Rikard, David E Bowen. 1989. Organization and customer: managing design and coordination of services. *Academy of Management Review* **14**(2) 213–233.

Nonaka, Ikujiro, Hirotaka Takeuchi. 1995. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press.

Park, Eun-Soo. 1995. Incentive contracting under limited liability. *Journal of Economics & Management Strategy* **4**(3) 477–90.

Paul, Alan K, Rahul N Anantharaman. 2003. Impact of people management practices on organizational performance: analysis of a causal model. *International Journal of Human Resource Management* **14**(7) 1246 – 1266.

Purcell, John, Nicholas Kinnie. 2007. Hrm and business performance. John Purcell Peter Boxall, Patrick M. Wright, eds., *The Oxford Handbook of Human Resource Management*. Oxford University Press.

Qian, Yingyi. 1994. Incentives and loss of control in an optimal hierarchy. *Review of Economic Studies* **61**(3) 527–44.

Redshaw, Bernard. 2000. Do we really understand coaching? how can we make it work better? *Industrial and Commercial Training* **32**(3) 106–109.

Wulf, Julie. 2007. Authority, risk, and performance incentives: Evidence from division manager positions inside firms. *The Journal of Industrial Economics* **55**(1) 169–196.

Wulf, Julie. 2012. The flattening firm: Not as advertised. *California Management Review* **55**(1) 5–23.