Internalizing Team Production Externalities through Delegation: The British Passenger Rail Sector as an Example

By RAFAEL HORTALA-VALLVE† and MIGUEL SANCHEZ VILLALBA‡
†London School of Economics ‡University of Alicante

Final version received 7 November 2008.

We show that in situations of team production, the associated externality between workers can be partially internalized via the delegation of authority of contracting rights, thus improving the overall efficiency of organizations. A delegated structure provides further incentives to individuals in higher tiers of the hierarchy because it makes them residual claimants. Our model sheds light on recent debates about the organization of the British passenger rail sector.

INTRODUCTION

When a Virgin train derailed near Graygigg (UK) on Friday 23 February 2007 and the sorrow associated with the casualties (including one death) settled down, old arguments regarding the organization of the British railway system resurfaced.

The agency in charge of the track (Network Rail) was found responsible of negligence, and several people (including Virgin chairman Sir Richard Branson) demanded a greater say in track maintenance by train companies. Britain’s biggest train company First Group and the Conservative party went further and proposed that train operators should do their own maintenance, thus going back to the pre-privatization scenario.

Following the Railways Act 1993, the state-owned British Rail was privatized. This implied a separation between maintenance (undertaken by Railtrack) and train operation in an attempt to improve the efficiency of the system. Nevertheless the industry’s safety record suffered (5 accidents, 59 deaths), so in 2002 the government created Network Rail (NR) to replace Railtrack. The latter’s reliance on subcontracted personnel was blamed for the crashes, so since its inception the new agency has used only its own staff for maintenance tasks.

The newly created Network Rail is monitored by the Office of Rail Regulation (ORR) but operates as a commercial business. Half of its income comes from the government and the rest is raised by access charges paid by the Train Operating Companies (TOC). The three actors (ORR, NR and TOC) can therefore be seen as the building blocks of a team production problem in which output is the quality of train services and the inputs are the efforts of the agents regarding the provision of the two basic determinants of quality, namely, track maintenance (NR) and train operation (TOC). These efforts are not verifiable, so contracts can be contingent only on output, i.e. on observable proxies for quality such as customer satisfaction or punctuality. The combination of joint production and contracts being contingent only on output, therefore, creates a positive externality between the agents: the effort exerted by one of them increases output, which in turn increases not only its own payoff but that of the other parties as well. Agents, however, fail to realize that such interaction exists and take into account only their own private benefits and costs when making their
investments, so that inefficiently low levels of effort are exerted. We show that the principal (ORR) can mitigate this problem by simply delegating contracting rights to one of the agents. In this way, the latter becomes the residual claimant and realizes the positive externality that its effort has on the other party. This means that the difference between private and social marginal benefits is reduced and that overall efficiency increases as a result.

In the absence of contractual restrictions, the Revelation Principle states that a delegated structure cannot improve upon a centralized one. In our setting, the principal could centralize the delegated structure by offering one of the agents a contract contingent on the subsequent contracts this agent writes with its subordinates. However, such contracts are not common in practice: rarely will a manager/subcontractor accept a contract that is contingent on their own actions as this will strip them of all freedom of choice. Indeed, no such contracts control the relationships between the component parties of the railway system, so it is under the assumption that those contracts are ruled out that we find that delegation is optimal.

The seminal reference on moral hazard in teams is Holmstrom (1982). Most of the contract theory literature on teams analyses the topic of collusion and the ways the principal has to avoid it (see Tirole 1986, and subsequent literature). In our model, however, collusion can only benefit the principal because it implies that the agents make decisions as a unit, so that they internalize the positive externality they exert on each other and increase their effort accordingly. This may yield a positive surplus to the agents if the principal does not anticipate such behaviour, but the principal’s profits would still increase relative to the situation where no collusion occurs. The literature has also analysed the role of hidden information and the effects of monitoring among agents (see Baliga and Sjostrom 1998; Faure-Grimaud et al. 2003). We abstract from such aspects in this paper.

In a similar vein to our work, Felli and Hortala-Vallve (2007) show how delegation can costlessly avoid collusion between a supervisor and an agent. Itoh (1991, 1994) focuses instead on the best way to allocate tasks among agents and whether the principal benefits from offering relative performance schemes. Finally, the closest references to our study are those of Macho-Stadler and Perez-Castrillo (1998) and Jelovac and Macho-Stadler (2002). They compare different contracting structures in the health sector using a binary effort model, but the key aspect of their work is the timing of events rather than delegation.

I. THE MODEL

The model presented below is based on Holmstrom and Milgrom (1987).

A Principal hires two identical agents \(i = 1, 2\) to undertake the production of a joint output \(x\). Agents are assumed identical in order to concentrate solely on the effects of different contracting structures. Output is assumed to be normally distributed with mean \(\mu\) and variance \(\sigma^2 > 0\). The expected output of the project increases in the effort exerted by the agents: \(\mu(e_1, e_2) = e_1 + e_2\). Production is sequential: first Agent 1 chooses her level of effort \(e_1\), then Agent 2 (after observing Agent 1’s choice) chooses hers. Agents participate only if their expected utility is no lower than their reservation utility (denoted \(U_0\)).

Efforts are non-verifiable and contracts can be contingent only on realized output \(w_i(x), i = 1, 2\). Moreover, we constrain such contracts to be linear in output, i.e.
Contract offers are assumed to be public (i.e. observable) by every party. We assume that the Principal can credibly commit to her proposed policy, thus avoiding the issue of renegotiation. Hereafter, capital letters denote aggregate variables: \( E = e_1 + e_2 \), \( W(x):= w_1(x) + w_2(x) \), etc.

The Principal is risk-neutral and maximizes the expected output minus total wages.\(^4\) Agents are assumed to be risk-averse with constant absolute risk aversion (CARA) utility and index of risk aversion \( r > 0 \). The disutility of effort is \( \frac{1}{2} e_i^2 \). The setting just described allows us to rewrite agents’ expected utility functions in terms of their certainty equivalent: \( a_i + b_i \cdot E - \frac{1}{2} b_i^2 - \frac{1}{2} e_i^2 \), where \( \gamma := r \sigma^2 > 0 \).

II. ALTERNATIVE ORGANIZATIONAL SETUPS

In a first best situation where efforts are contractible, the levels of effort exerted by the agents are \( e_1^* = e_2^* = 1 \) and the wages are \( w_i^*(x) = \frac{1}{2} + U_0 \), \( i = 1, 2 \). As expected, risk-averse agents face no risk and (since they are identical to each other) are treated identically.

The second best situation requires providing output-based incentives and insuring the agents against the resultant risk. This means that when the Principal optimizes, she needs not only to ensure that both agents’ participation constraints are satisfied, but also to take into account that they will react optimally to the contract they are offered (Incentive Constraints).

Depending on which actor can offer a contract to Agent 2, the second best scenario can be divided into two subcases: the Centralized Second Best if the offer is made by the Principal, and the Delegated Second Best if the offer is made by Agent 1.

Centralized Second Best (CSB)

We call Centralized Second Best the case in which the Principal offers both contracts: \( w_1(x) \) to Agent 1 and \( w_2(x) \) to Agent 2.

The timing of the game is as follows.

1. The Principal offers contracts \( w_1(x) \) and \( w_2(x) \) to Agents 1 and 2, respectively.
2. Agent 1 chooses her effort \( e_1 \) and exerts effort if her expected utility is above \( U_0 \).
3. After observing Agent 1’s participation and effort choices, Agent 2 chooses her effort \( e_2 \) and exerts effort if her expected utility is above \( U_0 \).
4. Output is realized and payments are made.

Therefore the Principal’s programme reads as follows:

\[
\max_{\{w_1(x), e_1, w_2(x), e_2\}} \mathbb{E}_x \{ \pi(x) \}
\]

\[
\left\{ \begin{array}{l}
\mathbb{E}_x \{ U(w_1(x), e_1) \} \geq U_0, \\
e_1 \in \arg \max_{\hat{e}_1} \mathbb{E}_x \{ U(w_1(x), \hat{e}_1) \}, \\
\mathbb{E}_x \{ U(w_2(x), e_2)|e_1 \} \geq U_0, \\
e_2 \in \arg \max_{\hat{e}_2} \mathbb{E}_x \{ U(w_2(x), \hat{e}_2)|e_1 \},
\end{array} \right.
\]

\( \text{II. ALTERNATIVE ORGANIZATIONAL SETUPS} \)

In a first best situation where efforts are contractible, the levels of effort exerted by the agents are \( e_1^* = e_2^* = 1 \) and the wages are \( w_i^*(x) = \frac{1}{2} + U_0 \), \( i = 1, 2 \). As expected, risk-averse agents face no risk and (since they are identical to each other) are treated identically.

The second best situation requires providing output-based incentives and insuring the agents against the resultant risk. This means that when the Principal optimizes, she needs not only to ensure that both agents’ participation constraints are satisfied, but also to take into account that they will react optimally to the contract they are offered (Incentive Constraints).

Depending on which actor can offer a contract to Agent 2, the second best scenario can be divided into two subcases: the Centralized Second Best if the offer is made by the Principal, and the Delegated Second Best if the offer is made by Agent 1.

Centralized Second Best (CSB)

We call Centralized Second Best the case in which the Principal offers both contracts: \( w_1(x) \) to Agent 1 and \( w_2(x) \) to Agent 2.

The timing of the game is as follows.

1. The Principal offers contracts \( w_1(x) \) and \( w_2(x) \) to Agents 1 and 2, respectively.
2. Agent 1 chooses her effort \( e_1 \) and exerts effort if her expected utility is above \( U_0 \).
3. After observing Agent 1’s participation and effort choices, Agent 2 chooses her effort \( e_2 \) and exerts effort if her expected utility is above \( U_0 \).
4. Output is realized and payments are made.

Therefore the Principal’s programme reads as follows:

\[
\max_{\{w_1(x), e_1, w_2(x), e_2\}} \mathbb{E}_x \{ \pi(x) \}
\]

\[
\left\{ \begin{array}{l}
\mathbb{E}_x \{ U(w_1(x), e_1) \} \geq U_0, \\
e_1 \in \arg \max_{\hat{e}_1} \mathbb{E}_x \{ U(w_1(x), \hat{e}_1) \}, \\
\mathbb{E}_x \{ U(w_2(x), e_2)|e_1 \} \geq U_0, \\
e_2 \in \arg \max_{\hat{e}_2} \mathbb{E}_x \{ U(w_2(x), \hat{e}_2)|e_1 \},
\end{array} \right.
\]
where the last two constraints are Agent 2’s Participation and Incentive Compatibility constraints and highlight the fact that Agent 2 observes Agent 1’s effort choice at the time of choosing her own level of effort $e_2$.

The second best efforts and contracts are

$$e_1^{**} = e_2^{**} = \frac{1}{1 + \gamma}$$

and

$$w_i^{**}(x) = a^{**} + \frac{1}{1 + \gamma} \cdot x,$$

$i = 1, 2$, respectively, where $a^{**}$ makes the participation constraints binding.

Moving from the single-agent to the multi-agent case, a positive externality between the agents enters the picture: when an agent increases her effort, she increases the expected wages of all the agents. Due to contractual restrictions, the Principal is unable to induce agents to internalize this externality when contracting with them in a centralized way. As a matter of fact, our simple model tells us that the best possible arrangement involves integrating all the activities so that effort decisions are made jointly. However, this is not possible in the case of the British railway industry because unmodelled aspects of it prevent such arrangements being reached, especially the EU Directive 91/440 that requires all EU member states to separate ‘. . . the management of railway operation and infrastructure from the provision of railway transport services, separation of accounts being compulsory and organisational or institutional separation being optional’.

**Delegated Second Best (DSB)**

The Principal can improve on the Centralized Second Best by changing the contracting structure to establish a hierarchy between the agents. The rationale for the improvement derives from the fact that the agent higher in the hierarchy recognizes the positive externality that her effort has on the other agent and is willing to exert more effort.

Under the delegated structure, the Principal contracts with one agent (Agent 1), who then subcontracts with the remaining one (Agent 2). The timing of the game is as follows:

1. The Principal offers contract $W(x) = A + Bx$ to Agent 1.

2. Agent 1 simultaneously chooses her effort $e_1$ and the contract she offers to Agent 2, $w_2(x) = a_2 + b_2x$; Agent 1 exerts effort if her expected utility is above $U_0$.

3. After observing Agent 1’s participation and effort choices, Agent 2 chooses her effort $e_2$ and exerts effort if her expected utility is above $U_0$.

4. Output is realized and payments are made.

The programme that the Principal faces reads as follows:

$$\max_{\{W(x), e_1, w_2(x), e_2\}} E_x\{V(x)\}$$

s.t.

$$\max_{\{e_1, w_2(x), e_2\}} E_x\{U(W(x) - w_2(x), e_1)\} \geq U_0,$$

$$E_x\{U(W(x) - w_2(x), e_1)\} \geq U_0,$$

s.t.

$$e_2 \in \arg \max_{e_2} E_x\{U(w_2(x), e_2)|e_1\}.$$
constraint of Agent 2 is binding. Since Agent 2’s participation constraint enters the Lagrangian for Agent 1’s programme directly, it is clear that Agent 1 now internalizes the effect of her effort on Agent 2’s output:

\[ \mathcal{L} = (A - a_2) + (B - b_2) \cdot E - \frac{1}{2} (B - b_2)^2 - \frac{1}{2} e_1^2 + \lambda \left( a_2 + b_2 \cdot E - \frac{1}{2} b_2^2 - \frac{1}{2} e_2^2 \right) + \varphi (e_2 - b_2). \]

The derivative of \( \mathcal{L} \) with respect to \( e_1 \) should be equal to zero at the optimum:

\[ \frac{\partial \mathcal{L}}{\partial e_1} = (B - b_2) - e_1 + \lambda \cdot b_2 = 0. \]

The above condition differs from the incentive constraint under the centralized structure because it includes the term \((\lambda \cdot b_2)\). This term captures precisely the positive externality that Agent 1 has on Agent 2: more effort by Agent 1 increases the expected output, which in turn relaxes the participation constraint of Agent 2.

The unique solution has all participation constraints binding and the following effort levels:

\[ e_1^d = b_2^d = \frac{2 + 3\gamma}{(1 + \gamma)(2 + \gamma)}, \]

\[ e_2^d = b_2^d = \frac{2 + 3\gamma}{(1 + 2\gamma)(2 + \gamma)}. \]

### III. Results

**Proposition 1.** The delegated structure always yields a higher expected profit to the Principal than the centralized one. The relative gains are higher, the higher the risk of the project and/or the higher the index of risk aversion of the agents.

The proof is immediate from comparing

\[ \Pi^{**} = \frac{1}{1 + \gamma} - 2U_0 \]

and

\[ \Pi^d = \frac{1}{2} \left( \frac{(2 + 3\gamma)^2}{(1 + 2\gamma)(2 + 3\gamma + \gamma^2)} - 2U_0 \right) \]

and showing that \( \Pi^d / \Pi^{**} \) is increasing in \( \gamma \).

An important corollary of this result is that the Principal cannot replicate the DSB outcome by offering the DSB contracts in a centralized fashion. This can be easily proven using a revealed preference argument: when contracting in a centralized way, the Principal can choose any pair of contracts \((\tilde{a}_1, \tilde{b}_1), (\tilde{a}_2, \tilde{b}_2)\), including those corresponding to the delegated structure, \((a_1^d, b_1^d), (a_2^d, b_2^d)\); in spite of this, she chooses the CSB contracts \((a_1^{**}, b_1^{**}), (a_2^{**}, b_2^{**})\).
Before moving to the next proposition, remember that the disutility of risk borne by Agent $i$ is equal to $\frac{1}{2} \text{Var}(w_i(x))$. Thus Agent 2’s disutility of risk is $\frac{1}{2} \left(b_2^d\right)^2$, and Agent 1’s is $\frac{1}{2} \left(B^d - b_2^d\right)^2$.

**Proposition 2.** Under the delegated structure, Agent 1 exerts more effort than Agent 2. Moreover, both agents exert more effort than under the second best structure, i.e. $e_1^d > e_1^{**} = e_2^{**}$ for all $\gamma > 0$. In terms of risk, Agent 1 bears less risk than Agent 2 ($b_2^d > B^d - b_2^d$) and the latter bears more risk under the delegated structure than under the centralized one ($b_2^d > b^{**}$).

Since both agents exert more effort under delegation than under centralization, expected output is greater under delegation. We also find that managers (agents higher in the hierarchy) exert more effort than subordinates, as is the case in Prendergast (1995).

From Agent 1’s programme we find that
\[
\frac{\partial b_2^d}{\partial B^d} = \frac{1 + \gamma}{1 + 2\gamma} \in \left(1, \frac{1}{2}\right);
\]
i.e. whenever the Principal induces a higher effort on Agent 1, the latter also provides more incentives to Agent 2. From the Principal’s perspective this generates a (second-order) trickle down effect that multiplies the initial (first-order) effect of an increase in $B^d$ by increasing also $b_2^d$. This effect is decreasing in $\gamma$.

Notice that the slopes of the wage contracts, besides providing information on the optimal levels of effort, are also related to the risk borne by the agents. Moreover, while Agent 1’s effort depends on her gross wage ($e_1^d = B^d$), her disutility of risk $\frac{1}{2} (B^d - b_2^d)^2$ depends on her net wage instead. As a consequence, Agent 1 is able to transfer part of her risk to Agent 2, though at the expense of exerting more effort than the latter. This result also shows how the fundamental trade-off in moral hazard situations, that of incentives versus risk, is lessened thanks to the delegation of contracting rights: for Agent 1, more incentives (greater $B^d$) does not mean as much extra risk ($B^d - b_2^d$) as in the centralized case since, as mentioned above, she will transfer some of the risk to Agent 2 by increasing $b_2^d$.

At first sight, the fact that the agent who is higher in the hierarchy bears less risk than her subordinates may seem puzzling. However, it is well known that outsourcing, subcontracting or decentralization are ways to pass some risk to those at the lower levels of the hierarchy. Indeed, their tendency to generate precarious, deregulated working conditions is one of the main criticisms levelled at practices such as subcontracting in the construction industry, outsourcing in manufacturing and the privatization of public services.

### IV. Discussion

In spite of the convexity of the disutility of effort and the fact that agents are homogeneous and risk-averse, the delegated structure distributes asymmetrically the risk and the effort. However, the associated internalization of the externality overcomes the inefficiency generated by the unequal treatment of agents, and overall efficiency is higher than under the centralized structure.

Two modelling assumptions are needed for our results to hold. First, each agent’s effort choice depends only on the power of her own gross incentive scheme (slope of her
contract) and is independent of the effort exerted by the other agent (i.e. there are no strategic complementarities or substitutabilities). Consequently, any agent affects the other only by modifying the latter’s participation constraint. Second, under the delegated structure Agent 1 fully internalizes her externality on Agent 2, hence her effort depends not on the distribution of the wages among the agents but only on the aggregate wage bill \( W(x) \).

When we consider a more general setting, the analysis becomes ambiguous precisely because the previous two observations no longer hold. For instance, strategic complementarities reinforce the pre-eminence of the delegated structure over the centralized one, while strategic substitutabilities work against the incentives of Agent 1 to exert more effort under the delegated structure.

The benefits of the delegated structure stem from the fact that Agent 1 internalizes the externality. This happens when she takes advantage of the interaction of her two choice variables \( e_1 \) and \( w_2(x) \); i.e. the individual at the top of the hierarchy not only proposes a contract to her subordinates but also decides her plans for the future (her effort decision). It would be boundedly rational for her not to take into account the interaction between both decisions. Indeed, in the British railway example, we can expect that if the Office of Rail Regulation (Principal) contracts with, for example, Network Rail (Agent 1), then the latter will choose simultaneously its investment in track maintenance and the access charges that the Train Operation Companies (Agent 2) have to pay in order to use the tracks.

Given the nature of both agents’ activities, our model seems to suggest that ORR should delegate contracting rights to NR, who should in turn contract with the TOC (note that our model and results naturally extend to the situation where there are various TOCs). The rationale for this suggestion is that, since the same piece of railtrack may be used by several TOCs, the alternative delegated structure (i.e. that in which TOCs are given joint control over it) will imply the multiplication of fixed costs and/or costly coordination between companies. These problems would be eliminated if control is given to a unique maintenance company (NR), which would also be allowed to choose the fees that the TOCs should pay in order to use the track.

Finally, we remain silent regarding the debate on the public or private ownership of any of the firms. Our analysis was kept within the present regulatory regime induced by the European Union directive 91/440 and, given this restriction, we proposed a contractual framework under which passenger railways services could be improved.

V. CONCLUSION

We analysed how team production must be organized when contracts can be written contingent only on joint output. We find that the positive externality associated with this setup can be partially internalized by the delegation of contracting rights, which decreases the misalignment between the principal’s and the agents’ interests and thus increases the overall efficiency of the organization.

Our policy conclusion regarding the organization of the British passenger rail services suggests that Network Rail should be allowed to choose how to allocate its investment in track maintenance as well as to determine (without direct interference from the Office of Rail Regulation) the access charges that train companies should pay for using the network.
ACKNOWLEDGMENTS

We thank Jordi Blanes, Frank Cowell, Matthias Dahm, Leonardo Felli, two anonymous referees, and seminar participants at various conferences for helpful comments and discussions. The first author gratefully acknowledges financial support from the British Academy. The second author gratefully acknowledges financial support from the Proyecto SEJ 2007-62656 (Ministerio de Economía de España). A longer version of this paper was previously circulated as ‘Hierarchic contracting’.

NOTES

1. Alternatively, as is common in the moral hazard literature, efforts can be considered to be verifiable and verification to be costly, a reasonable assumption in this case, where it is necessary to monitor a vast network.
2. If tracks are not properly maintained, trains need to slow down and punctuality may suffer. Conversely, if train companies do not instruct their drivers or maintain their trains properly, they cannot take advantage of well-maintained tracks. Efforts could be interpreted as maintenance or training activities, but one could also think of efforts as investments, provided that they are difficult (or too costly) to verify in a court of law and hence no contracts can be written upon them.
3. In our setting, whenever contracts are not constrained to be linear, and arbitrarily large punishments are allowed, first best can be arbitrarily approximated (Mirrlees 1999). However, this scenario is not realistic within our setting of passenger rail services.
4. Allowing for a risk-averse principal does not affect our qualitative results.

REFERENCES


© The London School of Economics and Political Science 2009