~~When markets do not work, other institutions may be created that do a better job. In particular, the firm can be such a mechanism.~~

ROBERTS, J. (2007), *The Modern Firm: Organizational Design for Performance and Growth*, 1st ed., Oxford University Press, Oxford, 88-103.

## Firms versus Markets

When might firms be better than markets? Much of economists' current understanding of the answer to this question traces to another element of the work of Ronald Coase. Almost 70 years ago, Coase (1937) asked explicitly why some economic activity is carried out through market transactions while other parts are organized under hierarchic authority relations within firms. His answer is that there are costs to organizing economic activity, to achieving coordination and motivation, and that economizing on these *transaction costs* explains the patterns of organization that are adopted. In particular, a transaction is removed from the arm's length contracting of the market and brought inside the firm precisely when it is cheaper to organize it this way. Thus, we are to understand the boundaries of the firm and, more generally, observed patterns of organizational design as being efficient ones—ones that create the most possible value.

There are at least two aspects of Coase's answer that need elaboration. One is why efficiency—rather than, say, the pursuit of monopoly power and profits—should be determinative. The second is the origin and nature of transaction costs.[8]

The basis for efficiency arguments is simply that if arrangements are not efficient, then (definitionally) it is possible to make everyone better off—not just to increase

the size of the total pie, but instead actually to increase the size of each person's slice. Presuming that the potential improvements can be identified and the gains can be shared, we should expect such changes to be made. So if an arrangement persists, there is reason to suspect that it is efficient, at least for the parties who are in a position to have their interests represented.[9]

This argument might seem Panglossian—that everything we observe is the best it could be—but there are important qualifications and subtleties that render it less problematic.

Note first that the requirement that the relevant parties can identify any potential improvements limits what is achievable. Thus, the efficiency of actual arrangements is constrained by informational and observational limitations.

An example is the "Market for Lemons" problem of adverse selection considered earlier. The result of the informational asymmetry may be that only the very worst cars are offered and sold, even though there are many potential trades that would make both sides to the transaction better off. Gains from trade are not realized because informational asymmetries prevent identifying their magnitude and sharing them in a satisfactory way. So this situation may, in fact, be the best that can be achieved (provided we continue to respect private property and allow each side to decide whether it wants to trade). Thus, it is efficient in the limited sense that we use the term, although this mainly demonstrates how weak a notion efficiency actually is, or just how constraining the informational limitations are.

Similarly, strikes and other costly delays in reaching agreements need not be seen as inefficient waste and

evidence against the efficiency hypothesis. Rather, they may be interpreted as the best options available under the circumstances for credibly communicating the value of an agreement to each side (Kennan and Wilson 1993). For example, a firm's willingness to suffer a strike signals that an agreement is not so valuable to it as might have been believed. Thus, the union learns that it cannot hope for as rich a settlement as it had desired. For if an agreement were really very important and valuable to the firm, it would be anxious to settle.

We now turn to the second point: the nature of transaction costs. In a market setting, transaction costs are the costs of finding and qualifying trading partners, of establishing specifications and prices, of negotiating and drafting contracts, and of monitoring and enforcing agreements. They are also the opportunity costs of lost benefits that are occasioned by the difficulties of developing complete, enforceable agreements between separate parties.

To a large extent, the informational and commitment issues discussed already underlie the transaction costs of using markets. However, one particular example has a central place in the research in this area. The example involves hold-up and specialized investments (Williamson 1975, 1985; Klein, Crawford, and Alchian 1978).

Williamson (1975) argues that many business dealings involve lock-in—even if there are initially lots of potential trading partners, once one is chosen and the parties start to work together, there is a "fundamental transformation" of the relationship that makes changing to another partner very difficult. In such circumstances, if contracts are incomplete, then the parties may have to negotiate after

the lock-in has occurred. These negotiations may be costly and acrimonious in the best of circumstances. Moreover, they present an opportunity for one party to act opportunistically to attempt to extract more of the returns to cooperation than it was due under the original agreement. Both the bargaining costs and potential benefits that are lost if the bargaining breaks down and cooperation does not occur are transactions costs of dealing with another party.

Lock-in is actually inevitable when assets are specialized. An asset is specialized to a particular use when the value it can create in its next-best alternative use is substantially lower than what it yields in the current one. For example, the dies used to shape materials in manufacturing are very specific to that use: If they are not employed for this purpose, they are just scrap metal. Firm-specific human capital—knowledge that is only (or especially) valuable in the context of employment with a particular firm—is another example. When assets are specialized, they are subject to hold-up—attempts by trading partners to appropriate some of the returns that the assets' owners expected when they invested in them. This can lead to a variety of inefficiencies.

Suppose two firms have an opportunity to trade, but the seller needs to make specific investments to serve the buyer's needs in the best way possible. Once the investments are made, the costs are sunk. This means that even if the price ultimately received by the seller were cut almost to the level of variable costs, so that almost no contribution to covering the costs of the investments would be realized, it would still not be worthwhile to withdraw from serving the buyer. The reason is that the sunk costs

must be borne in either event and the asset has no other good use. A portion of the returns to the asset are then *quasirents*, returns in excess of what is needed to keep the asset in its current use once it has been created.[10] The seller is then subject to the danger of hold-up.

If there is a prior contract, but it is sufficiently incomplete that negotiations over terms need to take place after the investments have been made, then the bargaining over terms will very likely give little protection to the seller's investments. This is because the sunk costs are irrelevant in determining how much value is created by cooperation versus breaking off the relationship, which is what the parties effectively bargain over. Even if the terms are nominally fixed in advance, the buyer may still be tempted to force a renegotiation of the terms of trade, appropriating a portion of the quasirents that the seller had hoped to enjoy. This is possible because the seller has little recourse: To refuse to renegotiate and break off the deal leaves him or her with only the nearly worthless asset. Meanwhile, forcing a renegotiation may be quite simple. For example, the buyer could claim business conditions have changed in way that justifies a lower price, or that service or quality has not been acceptable, or any of a number of other things, depending on the particular circumstances.

Thus, the seller cannot expect to receive the full returns on the specific investments it has made. Anticipating this, the setter may be reluctant to commit resources to the specific assets. For example, if the specificity arises from the seller's learning the particular needs of the buyer, the seller might underinvest in this knowledge, so that less of a loss is suffered in case of a hold-up. Thus, less value is

created. Alternatively, the seller may expend resources for protection against the anticipated hold-up. Making the assets more flexible in their uses, so that they can be redeployed at less cost, might do this. This is a waste, for the resources are being expended to improve the value of the asset in a use to which it ought not to be put.[11]

One solution to this problem is for the buyer to pay a part of the cost of the investment up front—essentially the buyer pays *ex ante* for the amount to be (mis)appropriated later. This will work, however, only if the agreement to undertake the investment is enforceable. Otherwise, for instance, the seller might just pocket the buyer's money and still make only the investment that seems individually optimal given that the terms will later be renegotiated. Another solution may be for the transaction to be brought within a single firm. Empirically, this has been an important element in vertical integration.[12] This is can be costly in a number of ways, however, as we will soon see. Thus, the enforceability problems create transaction costs in markets.

What are the transaction costs of organizing economic activities inside the firm? This is still a controversial issue. One might think first of the costs of communicating information up and down through the hierarchy, of information overload at the center/top, and of slow decision-making that is based on limited and possibly outdated information. Organizational decentralization may sometimes provide an effective response to these phenomena, however, as the developers of the multidivisional form discovered (Chandler 1977). (Generally, it is a good idea not to rely on explanations that are based on inefficiency—managers are awfully good at creating new and better ways to do business more efficiently!).

In this vein, Oliver Williamson (1985) has pointed to the policy of *selective intervention* as a response to any inherent disabilities of the centralized, hierarchic organization of a firm. The idea is to replicate the workings of the market within the firm whenever this yields efficiency, while top executives intervene selectively in the subunits and the relations among them only when this yields a better outcome than market dealings.

If selective intervention worked, then it would be efficient to have everything in one gigantic firm. Yet, even the ideologues of the old Soviet economy never dreamt of a system that was so extreme in its centralization. There must be something that prevents effective application of selective intervention.

One response is that it is impossible to generate the same intensity of incentives within a single integrated firm as when units are separately owned. In this regard, Williamson himself suggested that, while it might be easy to promise as strong incentives to employees as to outside contractors, it is hard to do so *credibly*. The problem is that the owner controls the performance measures[13] and would always be tempted to fudge them. This could happen both when the employee has done very well and is due to be paid a lot, and when there have been bad results despite apparently good effort, in which case the owner may be too forgiving. Either possibility blunts actual incentives and can imply that the firm does not achieve the levels of efficiency that the market might realize.

This argument clearly rests on the difficulties of effective contracting. Reputational concerns may help counter it. As well, it may be possible in some circumstances to use third-party monitoring and auditing. For example,

BP used "self-help" figures—essentially, improvements in earnings not resulting from changes in crude oil prices or exchange rates—in its performance pay. It then employed an outside auditor to attest to the accuracy of the self-help numbers it calculated. Equity carve-outs and tracking stocks may be an especially interesting possibility here. For example, Thermo-Electron Corporation sold stakes in its business units to the public explicitly to "outsource performance evaluation." Outside equity investors are strongly motivated to act as monitors because their own funds are on the line, and the stock prices they generate become low-cost, objective performance measures that may have more credibility and integrity than any internally generated ones.

The "property rights" approach to the theory of the firm, developed by Sanford Grossman and Oliver Hart (1986) and Hart and John Moore (1990),[14] suggests another reason why it may be harder to give strong incentives in a larger, integrated organization. This logic is most applicable when thinking about owner-managed firms. Suppose such a firm is selling to an industrial customer. If the relationship is severed, the owner of the upstream firm still owns the assets in his firm (machines, brand name, etc.) and can redeploy them as he sees fit. In contrast, suppose the customer owns the assets, with the upstream manager now an employee running a business unit corresponding to the original firm. Now if the relationship collapses, the manager does not get to keep the assets.

As Grossman, Hart, and Moore (GHM) argue, this difference affects the relative bargaining position of the two parties in dividing up the value created by their cooperation. (Assume that it is impossible to specify the division

of value contractually in advance, so that it must be determined by bargaining after the value has been realized.) Thus, the ownership of assets determines the payoffs the parties receive.

These payoffs in turn affect the strength of the incentives the parties have to undertake investments that are complementary with the assets of the firm/business unit, such as learning how to work with the assets more effectively or developing a brand that increases the value of the goods that the buyer produces using the services of the assets. Getting a larger share of the return motivates investing more to create greater returns. So, who owns the assets affects investment and thus the value created. If there are two separate firms interacting through the market, the supplier owns the assets and has strong incentives to invest, but the buyer's incentives are weak. If there is vertical integration and the buyer owns the assets, then the employee–manager has weak investment incentives, although the buyer has strong ones. With incomplete contracts it simply is not possible to give the same incentives to an employee as an owner receives.[15]

Note the importance of incomplete contracts to this theorizing. If binding agreements were possible, then the division of the value created could be set contractually to provide incentives, or, indeed, the investments themselves could be governed by contract. (In this, the GHM theory is like the hold-up analysis discussed earlier.) Then equally strong incentives could be given inside the firm as outside—ownership and the boundaries of the firm would not matter.

Bengt Holmström and Paul Milgrom (1991) have argued that the issue is not just offering incentives that are strong

enough, but offering ones that are appropriately *balanced*. The full details are in their model of multi-tasking in agency relationships, which is discussed in detail in the Chapter 4, but the theory rests on two observations. First, they note that typically there are multiple ways that someone can spend time, many of which might be of value to an employer. But if these activities compete for the person's attention,[16] then the incentives offered for different activities must be comparable. Otherwise, the person will focus disproportionate amounts of her effort on those things that are especially well compensated and ignore the others. The second observation is that providing strong financial incentives is costly if the person is risk-averse, because it loads extra risk into pay. Further, the cost is greater the more difficult it is to measure performance. This means that, other things being equal, tasks where performance is hard to measure should not be given as intense incentives as ones that are more accurately observed.

Suppose now that two activities are desired. Think of one as producing output, which is easily measured, implying that the costs of providing strong incentives (in terms of the risk that the person bears) are low. In isolation, this activity should then be given strong incentives. The other can be thought of as some form of investment, where effort is hard to measure accurately and in a timely fashion. For example, it is hard to determine precisely the change in the long-term value of a division occasioned by its manager's efforts and decisions. Providing strong incentives for this investment activity is very costly. This is because doing so makes her pay highly random, since it is not determined solely by the manager's actions but also by the other uncontrolled factors that affect measured

performance. The manager will have to be compensated for bearing this risk, so the costs ultimately are borne by the employer.

It is obviously desirable that the manager both increase current performance and undertake the right investments that increase long-term value. If, however, strong incentives are given for improving current costs and revenues and weak ones for investments (as might be optimal if there were no interaction among tasks), then problems arise. The manager is tempted to mortgage the future, ignoring good investments, and concentrate on getting current performance up, even if this reduces total value created. The solution must be to provide balanced incentives. There are two ways to do this.

One solution is to sell the operation to the manager, who then bears the long-term consequences of her investment choices as well as the current effects of improving performance. This may, in fact, be a factor in the management buyouts that first became prominent in the 1980s. The second is to treat the manager as a salaried employee, giving relatively weak incentives for both short- and long-term performances. (These incentives might be implicit and subjective, perhaps through the opportunities for promotion.) The first solution means that the manager receives as strong incentives for generating future returns as current ones. The second means that both sorts of effort get equally muted incentives. In either case, the incentives are appropriately balanced, and both activities get some attention (but less, of course, in the low-incentives, employment regime).

The key point for the present discussion, however, is that if the employing firm continues to own the investment

opportunities, the employee must be given weak incentives for other activities—weaker than what would be received as owner of a separate firm. Thus, the market solution cannot be replicated inside the firm.

A fourth approach to the issue of why selective intervention does not work questions whether senior management will—or whether they even can—limit their interventions to those that are efficiency enhancing (see Milgrom and Roberts 1988*b*, 1990*a*, *c*, 1992: 192–4 and 269–77, 1998). A defining characteristic of the firm is that its executives have the unchallenged legal right to intervene in lower-level operations and decisions, to direct that very specific actions be taken, and to enforce these directives. (Indeed, they must have this power if selective intervention is to occur.) In contrast, outsiders (even the courts or regulators) cannot easily make such detailed interventions. So moving an activity out of the market and into the firm increases the opportunity for interventions, including ones that are not efficiency enhancing.

Excessive or inappropriate interventions might come for a number of reasons. First, senior managers may be tempted to intervene when they should not because, after all, it is their job to manage. They may also be too impatient, intervening when they see that at lower levels people might not do the absolutely best thing. This is understandable—mistakes are being made, after all—but costly. The intervention destroys both the opportunities and the incentives for the lower levels to learn. It also undercuts their autonomy and the very real performance incentives that come from that (Aghion and Tirole 1997). The senior managers can also have an overblown estimate of their own abilities, not trusting others to take the appropriate

actions (i.e. those that they would take themselves). Finally, implicit bribery of various forms might lead them to intervene.

But even if the executives are scrupulously honest and superbly competent, there may still be excessive interventions. The problem is that lower-level people will care a lot about the decisions that the firm makes, and they will have every reason to attempt to influence the executives to intervene, making the decisions in the way they like. For example, someone in the organization has to be assigned to Paris, Texas, and someone to Paris, France. One can imagine that candidates would exert huge amounts of effort to affect this decision. Similarly, one person will be promoted and another not, or one division's investments will be funded and another's not. There will be strong incentives for the interested parties to try to influence these decisions, and not necessarily in directions that increase overall value creation. Further, to know when and how to intervene, the senior executives will have to rely on information from the potentially affected parties.

Among the techniques of influence are biasing information provision, misdirecting effort (e.g. towards building the case for your side rather than attending to ongoing responsibilities), politicking, and worse. Collectively, these *influence activities* have three sorts of costs. First, resources are directly expended on influencing decisions, even when all that is accomplished is to shift their distributional consequences (which means that no extra value is created). Note that such efforts will also call forth defensive expenditures from those who are threatened. The second is that, to the extent that the influence activities are successful, bad decisions may be made. The third

is that the firm may be led to change its organizational design from what would otherwise be ideal in order to control the influence activities.

There are a variety of methods that can be employed to limit influence activities. One is to limit communication between executives and lower-level people. This limits the opportunity for politicking and strategic information provision, although it brings an obvious cost that some useful information is not transmitted. The "three strikes and you're out" policy employed at ABB Asea Brown Boveri, the Swiss–Swedish electrical power equipment and industrial products company, is of this sort (Bartlett 1993). Two managers who disagreed could take their issues to a higher level for resolution, but only twice. If they did it a third time, one or both was replaced.

A second approach is to structure decision processes so they are less susceptible to influence. Firm adherence to bureaucratic, inflexible rules can be an example. If salaries are completely determined by seniority and job assignment, then there is no point in politicking for a raise. The airlines' policy of assigning cabin crews on the basis of seniority similarly minimizes influence opportunities. Promoting on objective measures of past performance rather than on the apparent (i.e. less objective and more manipulable) qualifications for the new post can be rationalized as reducing the incentives for influence (as well as motivating current performance). A very lean headquarters is one means to commit not to intervene too often—HQ simply lacks the resources to mess around in the lower levels' business. Executives may also attempt to establish a reputation for not intervening in order to deter attempts at inducing interventions. Note, however, that this may

require not intervening even when it might seem appropriate in a particular instance.

A third approach is to limit the distributional consequences of decisions, so individuals will have less at stake. For example, pay compression and uniformity of treatment, even when other factors argue for differentiation, have this effect. This may explain the pressures that are common in organizations to apply standard procedures even when differentiation and special treatment might seem appropriate—to do otherwise is to invite everyone to try to make the case that he qualifies as an exception.

A final method of controlling influence takes advantage of the fact that the boundaries of the firm set limits to such internal influence activities. Putting activities in separate firms limits influence activities. For example, different pay and promotion policies can be employed in connection with the different activities if they are in different companies, whereas differential treatment within a single firm might lead to huge amounts of politicking and influence activities. The reason the boundaries of the firm matter is that it is pointless to campaign with one's boss over assignment to another firm, but it may be quite reasonable to do so if the transfer is within the corporation. Similarly, one advantage of using outside suppliers rather than in-house ones is that trying to discipline or replace the in-house supplier for poor performance invites influence costs, as do the establishment and adjustment of transfer prices.

Once it is established that there are costs to internal organization, the boundaries of the firm are determined by the Coasian formula: Organize transactions internally if and only if the costs of doing so are lower

than organizing through markets. We know quite a bit about market organization. How then to think about internal organization? What characterizes a firm? Why and how are firms different from markets? When will the firm be the favored form?

## The Nature of the Firm

Many authors, including Ronald Coase (1937) and Herbert Simon (1951), have identified the essential nature of the firm as the reliance on hierarchic, authority relations to replace the inherent equality among participants that marks market dealings. When you join a firm, you accept the right of the executives and their delegates to direct your behavior, at least over a more-or-less commonly understood range of activities. Simon argued that this may be an efficient response to the impossibility of foreseeing and contracting on what tasks will need to be undertaken—the need for coordination—and to the impossibility (high costs) of bargaining anew each time there is a change in the required activities. It is not a perfect solution, because the boss will not have an automatic incentive to take account of the employees' interests in choosing how they spend their time. Still, it can be better than a rigid prespecification of activities, as under a simple market contract.

Others—most notably Armen Alchian and Harold Demsetz (1972) and Michael Jensen and William Meckling (1976)—have challenged this view. They argue that any appearance of authority in the firm is illusionary. For them, the relationship between employer and