

3

Bargaining costs, influence costs, and the organization of economic activity

PAUL MILGROM AND JOHN ROBERTS

This chapter is concerned with the economics of organization and management, a relatively new area of study that seeks to analyze the internal structure and workings of economic organizations, the division of activity among these organizations, and the management of relations between them through markets or other higher-level, encompassing organizations.

The dominant approach to this subject is transaction-cost economics, as introduced by Coase (1937, 1960) and developed by several others since, most notably Williamson (1975, 1985). The main tenet of Coase's theory is that economic activities tend to be organized efficiently – that is, so as to maximize the expected total wealth of the parties affected.¹ In this context, two sorts of costs are customarily identified – those of physical production and distribution and those of carrying out necessary exchanges. Because these are typically treated as distinct and separable, the efficiency hypothesis becomes one of transaction-cost minimization: The division of activities among firms and between a firm and the market is determined by whether a particular transaction is most efficiently conducted in a market setting or under centralized authority within a firm.²

This approach has two conceptual problems. First, the total costs a firm incurs cannot generally be expressed as the sum of production costs – depending only on the technology and the inputs used – and transaction costs – depending only on the way transactions are organized. In general, these two kinds of costs must be considered together;

We thank Yoram Barzel, Patrick Bolton, John Ferejohn, Victor Goldberg, Ed Lazear, Oliver Williamson, and the editors of this volume for their helpful comments. We also thank the National Science Foundation and the John Simon Guggenheim Foundation for their generous financial support. Much of the work reported here was done while Milgrom held a Ford Visiting Professorship at the University of California, Berkeley, during the 1986–87 academic year.

efficient organization is not simply a matter of minimizing transaction costs.³ Second, the general theory is too vague to be useful. If an existing institution or arrangement appears to be inefficient, one can always claim that it is simply because the observer has not recognized all the relevant transaction costs. To give the theory more power and to generate more specific predictions, recent developments of transaction-cost economics have focused on identifying the major components of transaction costs and how they affect the efficient form of organization.

Our principal purpose here is to add two elements to this theory. First, we will argue that the crucial costs associated with using markets to carry out transactions (rather than bringing them within a more complex, formal organization) are the costs of bargaining over short-term arrangements between independent economic agents. This accentuation of short-term bargaining costs contrasts with received theory (as presented by Williamson), which emphasizes asset specificity, uncertainty, and frequency of dealings as the key factors. Second, we identify certain costs attached to the centralized, discretionary decision-making power inherent in formal economic organizations such as firms. Particularly important among these are the costs of essentially political activity within the organization, which we call *influence costs*: the losses that arise from individuals within an organization seeking to influence its decisions for their private benefit (and from their perhaps succeeding in doing so) and from the organization's responding to control this behavior. These costs are an important disability of centralized control and help to explain why integrated internal organization does not always supplant market relations between independent entities.

The remainder of this introductory section discusses the firm's role in traditional economic theory, that theory's failure to treat issues of organization and management, and the need to address such issues. The section headed "Transaction-Cost Economics" articulates the received elements of the transaction-cost approach to the economics of organization. The section titled "Critique and Extension of the Received Theory" contains our criticisms of this approach, including our basic arguments for the centrality of bargaining and influence costs. The sections on "Bargaining Costs" and "Costs of Centralized Authority" explore each of these costs in more detail. The section on "Influence Costs in the Public Sector" briefly uses the logic of influence costs to examine some issues of government. The final section contains general conclusions.

Until recently, economists paid little attention to the internal workings of business firms and other economic organizations. In standard microeconomic models, a firm is simply a collection of possible production plans together with a rule for selecting among them. Typical

rules include profit or share-value maximization for firms run by the owners of capital, average-wage maximization for labor-managed firms, and surplus maximization for public enterprises. These models allow no explicit role for management activities – the processes by which firms generate and evaluate decision alternatives; formulate, implement, and monitor plans; coordinate distant branch stores, factories, and offices; balance the competing interests of employees, owners, customers, suppliers, and creditors; and motivate them all to work in the general interest. Although these models could include such activities, they are not explicitly represented.⁴

Why have economists clung for so long to such an incomplete account of economic organizations? Historically, economic theory's chief task has been to explain how market economies, with so little centralized direction, could have performed as well as they have. The performance of Western European and North American market economies over the centuries is, by world standards, nothing short of spectacular. Moreover, recent experience with economic development elsewhere in the world confirms the connection between the West's remarkable economic growth and the prevalence of market organization: Recall the success of the market economies of Japan and Singapore, or contrast the recent economic performance of Eastern and Western Europe, of mainland China and Taiwan, or of North and South Korea. The great economic puzzle that the sustained growth and development of market-oriented economies poses is not that firms and other centrally managed organizations can achieve order in their affairs but that markets, with little apparent planning or explicit coordination, can direct available resources to such good effect.

In the two centuries since Adam Smith's original explication of how markets might guide economic activity to serve the public interest, economists have dissected, analyzed, refined, and formalized the theory of markets controlled by impersonal forces – the invisible hand. But even as these economists worked, the economies around them were changing. No longer were firms mostly family affairs with bookkeeping and management operations done at night when the shop was closed. Continuous production processes and specialized equipment came into use, and the very visible hands of engineers, chemists, and professional managers came to control production activities. In the United States, as Chandler (1977) has explained, the growth of the railroads and the telegraph opened national markets and made large-scale factory production economical. This strained the capacity of local suppliers, and so required factory managers to plan more carefully. With this planning came both a greater opportunity and a greater need to consider alternative ways of organizing production. Should an automa-

bile manufacturer make or buy its headlamps, batteries, and body parts? Should it own a network of dealers selling directly to the public, contract with existing distribution companies, or sell to independently owned retailers? Purchasing supplies or hiring services in the market came to be just one organizational alternative for acquiring the necessary inputs for production, assembly, distribution, and sales.

What determines which inputs a firm will acquire by ordinary market exchange and which it will produce itself? What difference does it make whether a firm produces an input itself, has a regular supplier produce it, or buys it on the market from the lowest bidder? The second formulation of the question shifts attention subtly away from the mechanical details of how production is arranged toward a focus on how the relationships between those who carry out the successive stages of production are managed. It suggests that whether production is arranged internally or externally need not determine what equipment will be used or which people will do the work. "Internal" and "external" production are just terms to describe in a very partial way how productive relationships are to be managed.

Economists who study organizations have come to see the market as but one alternative for solving the management problem of coordinating the diverse activities and interests of consumers and firms. Markets can then be fairly evaluated only by comparing them to other means of solving the same problem. A full evaluation cannot be made until a unified theory of management processes has been developed. Without such a theory, economists' recommendations about such bread-and-butter economic questions as whether to regulate monopolies and whether public or private organizations should provide services like education, communication, transportation, and so forth must be regarded as tentative, at best. Economists can no longer ignore the economics of organization and management.

TRANSACTION-COST ECONOMICS

Coase (1937) created transaction-cost economics by shifting the focus on the firm from technological possibilities and the maximization of some market objective to transactions and the management of relationships. *Transact*, as an intransitive verb, means "to do business with; negotiate."⁵ *Transaction costs* encompass the costs of deciding, planning, arranging, and negotiating the actions to be taken and the terms of exchange when two or more parties do business; the costs of changing plans, renegotiating terms, and resolving disputes as changing circumstances may require; and the costs of ensuring that parties perform as agreed. Transaction costs also include any losses resulting from in-

efficient group decisions, plans, arrangements or agreements; inefficient responses to changing circumstances; and imperfect enforcement of agreements. In short, transaction costs include anything that affects the relative performance of different ways of organizing resources and production activities.

As indicated earlier, a central tenet of transaction-cost economics is that production in capitalist, profit-oriented economies will tend to be organized so as to economize on transaction costs. For example, inputs will tend to be acquired in the market rather than produced by the firm when the costs of market transactions are less than those of internal transactions. The tenet does not specify how this tendency to economize on transaction costs arises. Careful planning by especially competent management may sometimes be responsible,⁶ as may imitation of successful firms by less successful ones or the growth of efficiently organized firms and the collapse of inefficiently organized ones (Nelson and Winter 1982).

In its general form, the tenet is not an empirical hypothesis: It is too nebulous to be confronted directly with evidence. To make specific predictions from the theory, it is necessary to identify the costs characteristically associated with transacting business in different ways and to discover how circumstances cause these costs to vary.

Oliver Williamson (1985) has proposed one framework within which Coase's theory can be made more specific and operational.⁷ Williamson's theory is based on an analysis of the costs of contracting in business relationships. Contracts (explicit or implicit) govern a firm's relationships with its suppliers, employees, customers, creditors, and shareholders. A central premise in Williamson's theory (foreshadowed in Coase's own work) is that any contract that calls for the future delivery of a good or service, the future provision of capital, or the future performance of work must be incomplete. That is, a contract can never specify exactly what actions are to be taken and what payments are to be made in all possible future contingencies. There are several reasons for this. First, parties cannot perfectly anticipate all the possible contingencies that may affect their costs of performing as promised, or even their ability to do so. Second, even for circumstances that can be anticipated, it is often more economical to respond when the need arises rather than to plan in advance for every foreseeable contingency (Lindblom 1959). Third, writing unambiguous contracts is difficult because of the limitations of natural languages (Quine 1960). Drawing up contracts with too many fine distinctions may simply increase the likelihood that emerging events will fall into areas of ambiguity or overlap, leading to disagreements that will have to be resolved after the fact. Finally, enforceable contracts can be made contingent only on in-

formation that the parties will share and that courts can verify. They cannot be based on information that only one contracting party will have, even if that information would be necessary for efficient future decisions.

What are the consequences of this incompleteness of contracts? If planning and contracting were complete and costless activities, parties to a contract would, after their initial agreement, act as one. They would determine in advance and in detail the best possible actions for every contingency that might arise, and the contract would specify that those actions be taken and would provide incentives to do so. In reality, because planning and contracting consume real resources and because perfectly explicit and freely enforceable contracts cannot be written, the theory posits that contracting parties content themselves with an agreement that *frames* their relationship – that is, one that fixes general performance expectations, provides procedures to govern decision making in situations where the contract is not explicit, and outlines how to adjudicate disputes when they arise. The differences among simple market contracting, complex contracting, vertical integration, and other ways of organizing transactions lie primarily in the institutions they specify for governing the relationship when circumstances not foreseen in the contract arise.

For the transaction-cost theory to explain the great variety of contracting practices that actually exist, it must identify the critical dimensions that favor one form of contracting over another. According to Williamson (1985, p. 52), “the principal dimensions with respect to which transactions differ are asset specificity, uncertainty, and frequency. The first is the most important and most distinguishes transaction-cost economics from other treatments of economic organization, but the other two play significant roles.”

Asset specificity refers to the degree to which an asset’s value depends on the continuation of a particular relationship. Consider, for example, a firm that rents a computer system and invests in software and training for the employees who will use the system. If an identical or perfectly compatible computer cannot be rented or purchased from another source, the software and employee training are specialized assets because they would lose much of their value if the firm switched to another computer system. A supplier who acquires specialized dies or locates a plant near a customer’s remote factory has similarly invested in specialized assets. Klein, Crawford, and Alchian (1978) have dubbed the profits an investor stands to lose from terminating a particular business relationship “appropriable quasi-rents.”⁸ Logically, although quasi-rents may exist any time costs have already been sunk, appropriable quasi-rents exist precisely when there are specialized assets.⁹

For concreteness, let us suppose that a supplier invests in specialized assets. The supplier’s worry is that a customer might behave opportunistically – that is, might try to force a reduction in future prices, curtail purchases, make unreasonable quality demands, increase the variability of demand and the number of rush orders, or take other actions that would diminish the supplier’s margins. Note that none of these concerns would arise if the two had a complete, enforceable contract. Moreover, if the assets were not specialized, these threats would again not be cause for great concern: The supplier would be protected by the option to shift the assets to other uses in which they could command an equal return. However, specialized assets, by definition, cannot be shifted to other uses without loss, so the investor may be forced to accept reduced margins, leading to a substandard return on investment.

Indeed, it has frequently been argued that concerns that the buyer will appropriate quasi-rents may lead the supplier to invest too little in specialized assets.¹⁰ As an illustration of this, Klein, Crawford, and Alchian cite the case of Fisher Body. In the 1920s, Fisher refused to build plants adjacent to the General Motors plants that the company served. The authors argue that Fisher Body quite rightly feared that such plant sites would make the company vulnerable to General Motors’ subsequent attempts to force reductions in its margins.

Similarly, a buyer might enjoy quasi-rents that are subject to appropriation by a supplier, or both parties might earn appropriable quasi-rents from their assets. Generally, Klein, Crawford, and Alchian emphasize the importance of *co-specialized assets* – ones that are most valuable when used together. For example, an electricity generating plant at the mouth of a coal mine is co-specialized with the mine. Generally, when different parties own co-specialized assets, at least one party enjoys a flow of appropriable quasi-rents.

One apparent option to mitigate the problem of appropriation of quasi-rents is to make the contract’s price and other terms more explicit and rigid and to impose greater penalties for breach of contract. However, this solution is itself costly. Adding rigidity to a contract may reduce the parties’ flexibility in responding to future circumstances. Alternatively, if clauses are added to specify in advance more contingencies and the corresponding responses, direct contracting costs rise and the likelihood of ambiguity in the contract’s provisions increases.

Uncertainty about what circumstances will prevail when future actions must be taken is the primary factor that makes complete contracting impossible. Greater uncertainty about what future actions will be appropriate makes rigid contracts, which recognize few contingencies, more likely to lead to bad decisions; they are therefore more

costly. Flexible contracts, too, entail costs. They are, of necessity, open to different interpretations and thus to effective renegotiation. They therefore do little to reduce the risk that quasi-rents will be appropriated. In this context, Coase's hypothesis is that parties will normally agree on the contractual arrangements in which these costs are minimized.

If opportunities to appropriate quasi-rents from a particular specialized investment arise frequently, then contracting parties may find it economical to craft a specialized governance structure to deal with these temptations. Depending on the nature of the transaction, many alternative structures may be available, and these may vary greatly in their complexity and costs. The simplest are generally worded contracts. These are intended to be interpreted (by the courts, if necessary) in the event of a dispute, but the parties involved in such agreements rely primarily on each others' goodwill and business reputation, standard procedures, and their continuing business relationship to smooth out disagreements without extensive bargaining. When the specialized investments and associated appropriable quasi-rents are not large, as with arrangements to deliver standard commodities at an agreed price, simple contracts may be entirely adequate. In other situations, more careful planning or governance may be needed. Then contracts can be more detailed. For example, they may include price-escalator clauses and clauses indicating penalties for breach of contract or how to deal with specified contingencies.¹¹ They may specify procedures for selecting and using arbitrators or private judges to substitute for courtroom litigation. Firms can also merge¹² and give executives authority for making decisions. Highly detailed contracts and specialized procedures for making decisions and resolving disputes are expensive to write or design, but the costs of writing and designing are fixed costs that, once sunk, can be applied again and again to similar transactions. Hence, detailed contracts and specialized procedures are most cost-effective when similar transactions are frequently conducted.

As Williamson (1985) states them, the predictions of transaction-cost economics can be summarized as follows: In comparing business relationships that occur in the same legal environment and at the same time, governance structures will be most complex and most finely crafted for transactions with (1) the greatest value of appropriable quasi-rents, (2) the greatest uncertainty about performance conditions, and (3) the greatest frequency.¹³ Beyond these, we can add the prediction by Klein, Crawford, and Alchian that co-specialized assets will be co-owned, because co-specialization means that separate ownership exposes one or both parties to appropriation of quasi-rents while the as-

sets' long life means that the frequency condition for the efficiency of specialized governance is met.¹⁴ Finally, we have an observation made by Grossman and Hart (1986). If ownership rights in some assets are not transferrable (for example, an individual's human capital of knowledge, skills, connections, and so on) and if these assets are co-specialized with one or more other assets, then the relative degrees of co-specialization will determine ownership patterns.

CRITIQUE AND EXTENSION OF THE RECEIVED THEORY

Our principal intent in this section is to argue the importance of bargaining costs in market relations and to identify certain costs of centralized decision-making authority.

Received transaction-cost theory emphasizes the implications of the incompleteness of contracts that cover actions to be taken in the uncertain future. However, we will argue that this emphasis is somewhat misplaced. Instead, we will show that the key to evaluating the efficacy of market transactions is the costs of negotiating suitably detailed short-term contracts. If these costs were always zero, then organizing economic activity through market exchange would always be perfectly efficient. On the other hand, when the costs of negotiating periodic exchange agreements are sufficiently high, then regardless of other factors, such as the presence or absence of specialized assets, potentially important savings are to be realized by placing the activity under a central authority, which can quickly settle potentially costly disputes.

To understand these claims we must first understand what we mean by the terms "short-term" and "bargaining costs." When describing contracts "short-term" refers to a period short enough so that all the information that is relevant for current decisions is already available. Short-term contracts, by definition, do not specify how to act in the longer term as new circumstances arise. We interpret "bargaining costs" expansively, just as we did the term "transaction costs," to include all the costs associated with multilateral bargaining, competitive bidding, and other voluntary mechanisms for determining a mutually acceptable agreement. Bargaining costs include not only the wages paid to the bargainers¹⁵ or the opportunity costs of their time, but also the costs of monitoring and enforcing the agreement and any losses from failure to reach the most efficient agreement possible in the most efficient fashion.

With these definitions, having zero short-term bargaining costs means that the bargainers require negligible physical and human re-

sources to reach efficient short-term contracts. (A short-term contract is *efficient* if there is no other feasible *short-term* contract that both parties would prefer.) However, by definition, bargainers cannot commit themselves through a short-term contract to restrict their long-term behavior in any way, even though they may recognize the long-term impacts of their short-term decisions. For example, the parties to a short-term contract may agree on what investments in specialized assets to make this year and who will pay for these, but they cannot commit themselves to behave benignly next year toward the party who, having paid for the investment, has appropriable quasi-rents.

To establish the key role of bargaining costs, suppose that the costs of negotiating short-term contracts were zero. We consider a two-party relationship (such as between a supplier and a customer) for which efficient production demands that the supplier, the customer, or both invest in specialized assets. We assume that the parties meet the standard assumptions of the transaction-cost literature in that each is a risk-neutral, financially unconstrained, expected-wealth-maximizing¹⁶ bargainer. The two also share common beliefs about the relative likelihoods of various future contingencies and both are farsighted in the sense that they understand how their current actions and agreements will affect future bargaining opportunities and behavior. They are also opportunistic in the sense that their behavior at any time does not depend on past unbonded promises or on how past costs and benefits have been shared. Finally, we assume that contracts governing prices and behavior in the distant future are prohibitively costly to write because too many contingencies need to be evaluated and described (that is, there is too much uncertainty), but that contracts governing prices, bonus payments, and the actions to be taken in the near term, over which the relevant conditions are already known, are costless to write.

In general Williamsonian terms, the situation involves opportunistic behavior, imperfect long-term contracting, specialized assets, and uncertainty about the future. According to transaction-cost theorists, these conditions are sufficient to prevent a market arrangement based on a series of short-term contracts from yielding an efficient outcome. Nevertheless, we claim that if the costs of bargaining over short-term arrangements were zero – a condition that is apparently consistent with our other specifications – then the market outcome would be efficient. That is, the actions taken by the parties both in the short run and in the long run would in all contingencies be identical to those that would have been specified in the “ideal contract” – the efficient (possibly long-term and complete) contract the parties would sign if there were no restrictions at all on contracting.¹⁷

Before proving this proposition and explaining the argument supporting it and the defect in received theory, we should emphasize two points. First, given our assumptions of risk neutrality and common beliefs, the actions taken under an efficient contract do not depend on the bargaining power of the parties involved: Only the distribution of the fruits of the bargain depend on bargaining power. Conversely, because the parties are risk-neutral, if the actions they take coincide with those that would be specified in the ideal contract, then the arrangement is efficient, regardless of the payments made between the parties. Second, we do not claim that the inability to write complete contracts has no effect on the way the parties' share risks: By their very definition, incomplete contracts imply a limited capacity to make intertemporal or contingent transfers. What is unaffected is the set of actions the parties will eventually take and hence the agreement's efficiency.

To establish the proposition, we consider a two-period problem with two parties. (The extension to more parties and periods is straightforward.) Bargaining over first-period actions (x_1 and x_2) and first-period transfers (s_1 and s_2) is costless in the above sense, and, once the second period arrives, it will be costless to bargain over the actions (y_1 and y_2) and transfers (t_1 and t_2) in that period. However, at the first date, no binding agreements can be made about second-period actions and transfers. We assume that net transfers are zero: $t_1 + t_2 = s_1 + s_2 = 0$. Let $V_1(x_1, x_2, y_1, y_2, \mu, \nu, \pi)$ be the payoff (benefits less costs) accruing directly to the first party in the second period. This depends on the actions taken in each period, the resolution of any uncertainty before first-period bargaining (as indicated by μ), any uncertainty resolved after first-period bargaining but before second-period bargaining (ν), and any uncertainty not resolved until after second-period bargaining (π). Define $V_2(x_1, x_2, y_1, y_2, \mu, \nu, \pi)$ similarly as the direct second-period returns to the second party. The presence of x_1 and x_2 as arguments of V_1 and V_2 reflects the possibility that these decisions may be investments with long-term payoffs, and the presence of y_i as an argument of V_i allows for the possibility that the second-period returns (quasi-rents) may be subject to appropriation: The returns to j 's first-period actions depend on i 's second-period actions. Note that risk neutrality implies the absence of income effects, so first-period transfers do not affect second-period payoffs. Also, note that both μ and ν are known when second-period bargaining occurs and that x_1 and x_2 are already fixed at that point.

By hypothesis, the agreement reached at the second date, given the circumstances $C = (x_1, x_2, \mu, \nu)$ that prevail then, will be efficient; that is, y_1 and y_2 will be chosen to maximize expected total wealth ($E[V_1 + V_2 | C]$). Letting $W_i(C) = t_i + E[V_i | C]$ be the portion of expected

total wealth that accrues to bargainer i , we then have $W_1(C) + W_2(C) = \max_{(y_1, y_2)} E[V_1 + V_2 | C]$. (The transfers t_1 and t_2 will depend on the parties' relative bargaining strength, but the optimal, chosen actions $y_1(C)$ and $y_2(C)$ will not.)

At the first date, the parties, being farsighted, will correctly forecast the agreement that would be reached in any circumstances C in the second period. Thus, each will evaluate first-period agreements according to the utility functions $s_i + E[U_i(x_1, x_2, \mu, \nu) | \mu] + E[W_i(C) | \mu]$, where $U_i(x_1, x_2, \mu, \nu)$ is the first-period payoff net of transfers to i when the actions taken are x_1 and x_2 and the outcome of the uncertainty is given by μ and ν . Since the short-term agreement reached in the first period is, by hypothesis, efficient, it maximizes the sum of these two valuation functions. With common beliefs, this is equal to $E[U_1 + U_2 + W_1 + W_2 | \mu]$. Hence,

$$\begin{aligned} & \max_{x_1, x_2} E[U_1 + U_2 + W_1 + W_2 | \mu] \\ &= \max_{x_1, x_2} E[U_1 + U_2 + \max_{y_1, y_2} E[V_1 + V_2 | C] | \mu] \\ &= \max_{x_1, x_2, y_1(C), y_2(C)} E[U_1 + U_2 + V_1 + V_2 | \mu]. \end{aligned}$$

But the first expression is the wealth achieved under short-term contracting in the absence of bargaining costs, and the last expression is the wealth that would be achieved under an efficient long-term contract. Their equality means that full efficiency is realized in the absence of short-term bargaining costs.

To illustrate, consider the relationship between Fisher Body (the supplier) and General Motors (the customer) analyzed by Klein, Crawford, and Alchian. Suppose the relationship lasts for two periods. In the first period, the parties reach an agreement about plant site and design (investments in specialized assets, corresponding to x_1 and x_2) and about the share of the cost of constructing the plant each will bear. Such an agreement specifies only the immediate actions the parties will take and how they will be compensated for these. In the second period, the parties negotiate prices, possibly a fixed transfer payment, quality standards, and a delivery schedule (t_1 , t_2 , y_1 , and y_2) in full knowledge of the circumstances then prevailing (e.g., current model year body designs, demands for various models, the costs and availabilities of steel and substitute materials, and so on, modeled in our equations by μ and ν , as well as the previously made investments). By our assumption of costless bargaining, regardless of the first-period agreement, the

second-period agreement will be efficient given the conditions that prevail then.

Now consider what would happen if the parties were to agree in the initial period to make the efficient plant site and design decisions.¹⁸ Then, the actions taken in the second period would, in all circumstances, agree with those specified under the hypothetical ideal contract. We therefore conclude that the parties could sign a short-term contract in the first period that would lead to them making efficient decisions in both the first and second periods. Actually, by varying who pays for the initial investment in the plant, all distributions of the fruits of these efficient decisions can be attained. Any contract, therefore, that leads to inefficient decision making can be improved upon for both parties by some contract that leads to efficient decision making. Thus, if the costs of short-term bargaining were zero, the agreement reached would indeed lead to efficient actions.

What, then, was wrong with the argument advanced in the first section of this chapter? Why shouldn't the fear of opportunism by General Motors make Fisher Body unwilling to enter into the arrangement? The answer is that Fisher can be compensated for the risk by having General Motors bear part of the plant's cost. Why, then, shouldn't General Motors fear that Fisher will appropriate its quasi-rents? Because the agreement can call for General Motors to pay for only as much of the plant's earnings as it expects to appropriate in future negotiations. Threats of appropriation are simply distributional threats; they are not threats to efficient action as long as bargaining costs are zero. Among risk-neutral parties with common beliefs and no private information, distributional threats can be compensated by initial cash payments. The efficiency of market arrangements is limited only by the costs of negotiating efficient short-term contracts. This conclusion points to the central importance of bargaining costs in determining the efficiency of market transactions. We shall study the origins and determinants of bargaining costs in the next section.

The preceding analysis relied on the assumptions that all parties are risk-neutral and that they can contract for current actions without restriction. The first of these assumptions is not reasonable when contracting parties are individuals, and the second fails when current actions cannot be precisely monitored. Nevertheless, as Fudenberg, Holmstrom, and Milgrom (1990) have shown, the conclusion that short-term contracts are as good as long-term contracts when no bargaining costs are involved applies equally to situations involving risk-averse bargainers and imperfectly observed actions, provided contractual payments in each period can be made to depend on any infor-

mation obtained during the period and provided no new information about any period's actions arrives only in later periods.

Our other criticism of early transaction-cost theories concerns their relative silence regarding the source, nature, and magnitude of the costs incurred in nonmarket transactions. Indeed, despite the firm beliefs of many economists that markets often hold great advantages over nonmarket forms of organization,¹⁹ received transaction-cost theory leaves unclear why market transactions are *ever* to be preferred to nonmarket ones.

Identifying the costs of general nonmarket transactions is a task to be approached with great caution. As Chandler (1962) has documented, business organizations have changed substantially and repeatedly over the past century, and the disabilities (transaction costs) suffered by an older form of organization may be overcome by its replacement. Perhaps wisely, then, transaction-cost theorists for a long time were largely silent about the source and nature of the costs of centralized organization, although they were certainly aware of the problem.²⁰ Quite recently, however, Williamson (1985) and Grossman and Hart (1986) have addressed explicitly the disabilities of nonmarket organization.

Williamson's treatment of the question of "Why can't a large firm do everything a collection of smaller firms can do, and more?" employs a crucially important idea: the notion of *selective intervention*. Many of the arguments purporting to explain the limits of organization fail when confronted with the policy of replacing previously autonomous units with semiautonomous ones in whose operations and decisions central managers intervene only when uncoordinated or competitively oriented decisions are inefficient. Any adequate explanation of why all economic activity is not brought under central management must confront this possibility.

Grossman and Hart (and Hart and Moore 1988) attempt to deal with this problem with a unified theory that treats the costs and benefits of different forms of organization as being all of a single type. Specifically, they identify asset ownership, with the possession of residual control rights over the assets – that is, all rights to the disposition and use of the assets that are not either given away in explicit contracts or claimed by the state. Ownership of a firm is then solely an issue of who retains these residual control rights over the collection of physical assets that Grossman and Hart identify with the firm. Because contracting is necessarily incomplete, such residual rights must exist. Moreover, Grossman and Hart assume that contracts are so incomplete, even in the short term, that parties cannot commit themselves to current actions, so that the analysis we have given does not apply. Under these

conditions, the allocation of control rights affects the ability of the parties to appropriate one another's investments and to protect their own investments from appropriation. Thus, just as the costs of a transaction between two independent owner-managed firms arise because each owner-manager's decision making ignores how his or her actions may benefit the other firm's asset values, the cost of integrating two previously independent firms is that the manager who is no longer an asset owner will ignore how his or her actions affect the integrated firm's assets: He or she will no longer manage these assets efficiently.²¹

Williamson's treatment of these issues is also based on incentive arguments. He focuses on why "high-powered", marketlike incentives that replicate residual claimant status are not feasible within a centrally managed organization – that is, he focuses on why selective intervention is not in fact possible. His answer is based on the idea that difficulties of (verifiable) measurement give rise to two moral hazard problems. First, the assets of the acquired stage will not be carefully managed because the manager cannot truly be the residual claimant, given that observation of the manager's actions is imperfect (or, at least, not contemporaneous); that resignation is an option; and that mechanisms for conveying reputations are imperfect. Thus, as in Grossman and Hart, assets will be mismanaged. Second, the returns of the acquired stage will be subject to appropriation via manipulation of the transfer prices and other accounting constructs that the center controls and that are too complex to be subjected to complete contracting. This, too, destroys incentives for proper asset management at the acquired stage.

These arguments have much to recommend them, but their focus on physical assets is too narrow. In particular, Grossman and Hart specifically do not distinguish between an organization with paid employees and one that contracts for labor services with independent suppliers but that owns and retains title to the tools and other physical assets that workers use in production. But what of the many firms, such as computer software development, public accounting, management consulting, and legal services firms or, to a somewhat lesser extent, universities and sports teams, whose only significant assets are the working relationships among their employees? Either the theory is silent in such cases, or it suggests that such organizations should have no bounds on their efficient size because they have no significant assets of the type Grossman and Hart consider.

A second criticism of the Grossman-Hart approach recognizes that incentives are a function of income streams, not just of decision rights, and that residual decision rights do not totally determine income streams when decisions have multidimensional consequences that ex-

tend over many periods and are not immediately and perfectly observable. A more satisfactory theory would integrate both factors.²²

Despite these criticisms, we believe that these incentive arguments have substantial force. Nevertheless, these theories miss an important class of generally identifiable costs of internal organization that do not depend specifically on control of assets. In the section titled "Costs of Centralized Authority," we argue that the crucial distinguishing characteristic of a firm is not the pattern of asset ownership but the substitution of centralized authority for the relatively unfettered negotiations that characterize market transactions. And, we argue, the very existence of this centralized authority is incompatible with a thoroughgoing policy of efficient selective intervention. The authority to intervene inevitably implies the authority to intervene inefficiently. Yet such interventions, even if they are inefficient overall, can be highly beneficial for particular individuals and groups. Thus, either inefficient interventions will be made and resources will be expended to bring them about or to prevent them, or else the authority to intervene must be restricted. This implies that some efficient interventions must be foregone.

BARGAINING COSTS

What are the costs of bargaining? We have defined these to include the opportunity costs of bargainers' time, the costs of monitoring and enforcing an agreement, and any costly delays or failures to reach agreement when efficiency requires that parties cooperate. Our analysis in this section will focus on costly delays and failures to reach agreement. The idea comes easily to economists that when parties in a bargaining situation have all the relevant information, they will agree to an efficient bargain. Nash (1950, 1953) elevated this proposition to an axiom in deriving his famous bargaining solution, and Coase (1960) made it the linchpin of his theory of property rights. Buchanan and Tullock (1962) made the same point in connection with their argument that only costs – inefficiencies – of private bargaining can justify government provision of goods or services:

If the costs of organizing decisions voluntarily should be zero, all externalities would be eliminated by voluntary private behavior of individuals regardless of the initial structure of property rights. There would, in this case, be no rational basis for state or collective action beyond the initial minimum delineation of the power of individual disposition over resources. (pp. 47–48)

The evidence supporting this idea, however, is mixed.²³ When experimental subjects are asked to divide a sum of money, say ten dollars, they have little difficulty agreeing to split the sum equally without costly delays or disagreements. But when the thing to be divided is

more complicated, so that symmetry does not focus the bargainers' attention on an obvious solution, posturing, haggling, and disagreement is more likely, as each party seeks to create or stake out a reasonable-sounding position that yields a large share of the available rewards.

To get a better idea of how serious these coordination difficulties might be, we turn to the analysis of a bargaining by demands game introduced originally by Nash (1950, 1953). Suppose that two parties have one dollar to divide. We can interpret the dollar as the maximal wealth attainable from exchange between the two parties. For example, it might represent the value a potential buyer puts on an object that is worthless to its current owner. The rules of the bargaining game are as follows. Each of the two parties, *A* and *B*, makes a demand, *a* and *b*. If the demands are consistent with the available resources – that is, if $a + b$ does not exceed one dollar – then each party gets what it demanded. If the demands are inconsistent with available resources, both parties get a payoff of zero.

If the problem were presented in just this way, the parties would very likely each demand fifty cents, resulting in a 50–50 split. In the terms Schelling (1960) used, the 50–50 split is an obvious *focal point* – a way for the parties to coordinate their demands. However, most real bargaining situations have either no focal points on which to coordinate, or many possible ones, which is just as bad. What should we expect to happen then?

For a game-theoretic analysis, we may ask, what is the full set of noncooperative equilibrium outcomes of this demand game? These outcomes represent patterns of behavior that are consistent with the rational and well-informed pursuit of self-interest on both sides. The answer is that for any pair of positive numbers summing to one dollar or less, there is a Nash equilibrium (possibly in mixed strategies²⁴) of the demand game at which the players' expected payoffs are precisely those numbers.²⁵ In particular, there is a Nash equilibrium in which both bargainers demand the whole dollar and, as a result, both receive zero.

This game-theoretic analysis not only captures the familiar idea that the division of the gains from trade may be indeterminate under bilateral monopoly, it also shows that the actual magnitude of the total gains realized may be similarly indeterminate.²⁶ The bargainers may fail to agree on any efficient solution, and, indeed, the resources that rational parties may squander in jockeying for bargaining position can be as little as zero or as large as the entire potential gains from trade.

Remarkably, the introduction of a minimal amount of competition virtually eliminates the potential for such coordination failures in two-party bargaining. Suppose, for example, that the bargaining situation involves two suppliers and a buyer. In terms of our model, there are

now three parties to the bargaining – A, B, and C, who make demands a , b , and c . The demands are compatible if either $a + b$ or $a + c$ is less than one dollar. The rules of the game are as follows. If the buyer's demand is inconsistent with both suppliers' offers, no agreement is reached, and each party receives a payoff of zero. Otherwise, buyer A does business with the supplier making the smaller demand or randomizes if the suppliers' demands are equal. If the buyer and a seller make consistent demands, each receives the amount demanded, and the other supplier gets zero. Almost all the "equilibria" of this "auction" version of the demand game are efficient.²⁷ Moreover, just as in a competitive market, the buyer receives all the surplus at equilibrium.

Variations on this three-party demand game lead to the same conclusion. For example, suppose that if the demands are consistent, one party gets one dollar minus the other party's demands or, alternatively that the parties split the difference. In each of these games, essentially all of the equilibria lead to the efficient outcome, in which the buyer receives all the surplus. This is a natural result of bidding competition among the suppliers.

These demand games can be interpreted as models of a competitive supply market in isolation. When perfectly competitive suppliers must make simultaneous offers, competition among them reduces the scope for disagreement with the buyer, leading to efficient outcomes. (Clearly, competition among buyers has the same effect.) The two-party demand game, by contrast, illustrates the inefficiencies that may result with a single supplier and purchaser. Specialized assets tend to generate bilateral monopolies which are accompanied by struggles for rents and consequent bargaining inefficiencies. Thus, specialized assets *cause* bargaining costs, which may explain the predictive successes of received transaction-cost theory.²⁸

The first class of bargaining costs, then, are coordination failures. They arise in situations where individuals could adopt several different patterns of mutually consistent, self-interested behavior and where market institutions fail to ensure that only efficient patterns actually emerge. Both standard economic theory and transaction-cost theory have typically assumed that, with competitive supply conditions, market mechanisms overcome these coordination problems. The analysis offered in this chapter does not contradict that view. However, recent studies involving detailed models of market institutions for price and quantity determination raise serious doubts about this assumption when multiple goods are involved and more than two parties must agree in order to benefit from exchange (Roberts 1987). In such situations, even when competitive pressures lead to perfectly competitive prices, coordination problems may still be so severe that beneficial ex-

change completely collapses. Of course, a key task of management is coordinating actions within an organization, so the case in favor of internal organization is strengthened by recognizing the possibility of coordination failures even in a system of competitive markets.

Measurement (information acquisition) costs are a second source of bargaining inefficiencies. Barzel (1982) and Kenney and Klein (1983) emphasize these costs to explain specialized contracting practices and vertical integration. They provide the basis for what has emerged as a second line of transaction-cost analysis – the measurement costs branch, in parallel with the asset specificity branch on which we focused in the previous section of this chapter. The idea is that individuals operating under standard short-term contracts will expend socially excessive amounts of resources to determine the private benefits and costs of an agreement when only its total costs and benefits, and not their distribution, matter for efficiency.

As an example of how measurement costs affect market arrangements, consider the Central Selling Organization (CSO) of De Beers, which in 1980 supplied between 80 and 85 percent of the world market in diamonds.²⁹ Kenney and Klein (1983) describe the CSO's marketing practices as follows:

Each of the CSO's customers periodically informs the CSO of the kinds and quantities of diamonds it wishes to purchase. The CSO then assembles a single box (or "sight") of diamonds for the customer. Each box contains a number of folded, envelope-like packets called papers. The gems within each paper are similar and correspond to one of the CSO's classifications. The composition of any sight may differ slightly from that specified by the buyer because the supply of diamonds in each category is limited.

Once every five weeks, primarily at the CSO's offices in London, the diamond buyers are invited to inspect their sights. Each box is marked with the buyer's name and a price. A single box may carry a price of up to several million pounds. Each buyer examines his sight before deciding whether to buy. Each buyer may spend as long as he wishes, examining his sight to see that each stone is graded correctly (that is, fits the description marked on each parcel). *There is no negotiation over the price or composition of the sight* [emphasis added]. In rare cases where a buyer claims that a stone has been miscategorized by the CSO, and the sales staff agrees, the sight will be adjusted. If a buyer rejects the sight, he is offered no alternative box. Rejection is extremely rare, however, because buyers who reject the diamonds offered them are deleted from the list of invited customers.

Thus, stones (a) are sorted by De Beers into imperfectly homogeneous categories, (b) to be sold in preselected blocks, (c) to preselected buyers, (d) at non-negotiable prices, with (e) buyers' rejection of the sales offer leading to the withdrawal by De Beers of future invitations to purchase stones. (p. 502)

What accounts for these nonstandard practices? In an ordinary market the buyers and seller would evaluate and haggle over each stone or

group of stones. The evaluation process would waste an inordinate amount of resources, and the haggling might even prevent agreement. Each buyer would carefully inspect each rough stone to determine how to cut it to create the largest, most flawless, and most valuable diamond and would use that information to estimate the stone's value. To bargain effectively, the seller must be equally well-informed, but to be so would require a substantial nonproductive investment. If the buyer and seller fail to agree on a price, another buyer would have to make the same evaluation, which would result in a duplication of effort and a waste of resources.

Given De Beers's initial classification of its diamonds, there is little social gain from further refining the allocation of diamonds among buyers. In a traditional market arrangement, customers would evaluate some stones that they will never cut, and the seller, in self-defense, would examine stones more closely than it would otherwise need to do. The De Beers system minimizes these measurement costs, which are attendant to haggling over price, and so represents one possible efficient response.³⁰

Notice how the De Beers system moves away from markets and introduces an element of centralization. Haggling is eliminated and the CSO is given authority to allocate the diamonds subject to certain categorization rules. Buyers who refuse their sights thereby terminate their relationships with De Beers. This is analogous to the right employees of any business have when they are unhappy with their wages or jobs; they can quit.

Even the most casual review of markets suggests many circumstances in which presale product evaluation and negotiation by buyers would not help allocate goods more efficiently but would give buyers an edge in bargaining. In such circumstances, alternative arrangements that economize on these costs should be expected. Barzel (1982) uses this idea to explain fruit and vegetable packaging (which discourages product evaluation) and product warranties (which make careful product evaluations less valuable to the buyer, and so reduce measurement activities).³¹ Kenney and Klein (1983) use it to explain the packaging of diamonds and the block booking of movies (which prevents theater owners from picking and choosing among new releases and so economizes on measurement costs). The royalties paid to authors of books can be similarly explained. If fixed fees were paid to an author, competing publishers would incur excessive costs in estimating the book's market potential for fear of the "winner's curse," according to which they acquire rights only to those books whose market potential they have overestimated and that other publishers, who have better esti-

mates, spurn. Compensating authors with royalties alleviates the winner's curse by making publisher's payments depend on actual rather than estimated sales.³² Part of the costs of allowing speculators to trade in a commodity market is that their profits must compensate for their socially unproductive investments in the information that is so essential to them (Hirschleifer 1971). The fact that these last markets are auction markets with little explicit negotiation has little import for our argument.

In general, initial uncertainty about a good's quality coupled with the possibility of resolving this uncertainty at some cost leads bargainers to act on this possibility, thereby increasing the costs of market arrangements. Such diverse arrangements as vertical integration, product warranties, and nonstandard market arrangements may emerge as the parties attempt to economize on these costs.

A third source of bargaining costs, and the one most often emphasized in the recent theoretical literature,³³ is private information about preferences. Unless the parties' valuations of a good being traded are common knowledge, the parties may be delayed in reaching an agreement or may even fail to agree at all, because they may strategically misrepresent the good's value. By insisting, for example, that "it's worth only fifty dollars to me, and I won't pay a penny more," a buyer can hope to get a lower price even though his or her actual valuation of the good may in fact be far greater. But this may prevent trade when the seller's value is relatively high, even though it is less than the buyer's true value. Moreover, given uncertainty about whether trade is efficient, bargaining costs of this form are absolutely inevitable, regardless of the bargaining procedure used (Myerson and Satterthwaite 1983). However, little is presently known about the determinants of these costs.

The role of uncertainty in generating bargaining inefficiencies dovetails nicely with Williamson's analyses, whether the uncertainty is about quality, with both sides initially being symmetrically informed but expending resources to acquire nonproductive information, or about parameters such as individual valuations, where informational asymmetries are inherent.

Our analysis of the sources of bargaining costs has been tentative and preliminary. Yet, it has served more than one valuable purpose. It has reinforced the logic of transaction-cost theory, provided a unifying perspective from which to investigate two previously distinct branches of transaction-cost economics — one based on specialized assets and one on measurement costs — and pointed to a new agenda for bargaining theorists and experimenters.

COSTS OF CENTRALIZED AUTHORITY

Accounts of Western economic growth often emphasize the importance of decentralized economic control rights (North and Thomas 1973). As Rosenberg and Birdzell (1986, p. 24) have recently written:

We have emphasized the part played by innovation in Western growth. The decentralization of authority to make decisions about innovations, together with the resources to effectuate such decisions and to absorb the gains or losses resulting from them, merits similar emphasis as an explanation of Western innovation. This diffusion of authority was interwoven with the development of an essentially autonomous economic sector; with the widespread use of experiment to answer questions of technology, marketing, and organization for which answers could be found in no other way; and with the emergence of great diversity in the West's modes of organizing economic activity.

Thus, Western economic history suggests that centralization stifles innovation. Is this a generalizable proposition? Even if one agrees that guild, church, and feudal authorities squelched experimentation and innovation in medieval Europe and that China's mandarinates, Japan's feudal lords, and Islamic mullahs did the same in their own domains, the historical record does not show that a modern central planner, who has studied the lessons of history, cannot guide an economy to duplicate and improve upon the performance of market economies. Yet the belief that such centralized planning and control stifles innovation is widespread; it even won official credence in the Communist economies of the Soviet Union and Eastern Europe.

Why can't a centrally planned, consciously coordinated system always do at least as well as an unplanned, decentralized one? For many years scholars, failing to find an answer to this question, have boldly (and, we think wrongly) concluded that there is no answer. For example, in his presidential address to the American Economic Association, Frederick Taylor (1929) held that socialist economies can allocate goods as well as capitalist economies because they can duplicate those economies in all their desirable respects:

In the case of a socialist state, the proper method of determining what commodities should be produced would be in outline substantially the same as that just described [for capitalist economies]. That is, the correct general procedure would be this: (1) The state would ensure to the citizen a given money income and (2) the state would authorize the citizen to spend that income as he chose in buying commodities produced by the state – a procedure which would virtually authorize the citizen to dictate just what commodities the economic authorities of the state should produce.

Substantially the same puzzle arises in trying to explain why there are any limits to a firm's size and scope. Thus, economists have asked, "Why, if by organizing one can eliminate certain costs and in fact re-

duce the cost of production, are there any market transactions at all? Why is not all production carried out by one big firm?"³⁴ And, "Why can't a large firm do everything that a collection of small firms can do, and more?"³⁵

The form of these questions assumes that benign, costless, *selective interventions* of the type Williamson considered are possible. This requires a decision maker with the authority to intervene, the interest in doing so only when appropriate, and the ability to consider and reject interventions without distorting the behavior of others in the organization. We argue that these requirements realistically cannot be met.

We take the view that what most distinguishes any centralized organization is the authority and autonomy of its top decision makers or management – that is, their broad rights to intervene in lower-level decisions and the relative immunity of their decisions from intervention by others.³⁶ Increases in centralized authority carry with them increases in the discretionary power to intervene. This increased power necessarily has costs that are avoided in more decentralized contexts. From this perspective, the principles that guide a firm's decision whether to manufacture an input (centralized organization) or to buy it from an independent supplier (decentralized organization) can be applied equally well to evaluate the relative productive efficiency of capitalist and socialist economic systems.

Two kinds of costs generally accompany increases in discretionary centralized authority. Both have the same fundamental cause: The very existence of such authority makes possible its inappropriate use. The first kind arises because those with discretionary authority may misuse it directly, on their own initiative. The second arises because others in the organization may attempt to persuade or manipulate those with authority to use it excessively or inappropriately. Inappropriate interventions, the attempts to induce them, and the organization's efforts to control both – all generate costs of increased centralization.

The first source of the costs of centralized, discretionary authority is inappropriate interventions that occur because individuals with increased authority are unable or unwilling to resist interfering where or in ways that they should not. This may happen simply because the individuals feel an imperative to manage – that is, after all, what managers are paid to do! Business people often cite this imperative to intervene as a characteristic and a cost of government bureaucracies: Bureaucrats look for something to do, whether or not their intervention is likely to be helpful. Private managers are presumably not immune to this failure, let alone to believing that their interventions will be beneficial when they are actually unlikely to be. Another possible reason for inappropriate intervention is that individuals in authority

may have personal interests in decisions: Will the empty lot next to the apartment building owned by the park commissioner's cousin be converted into a city park? Will the executive's protégé be appointed to replace a retiring division head? For any of these reasons, authority will be exercised more often and in other ways than efficiency alone dictates.

In a related vein are the costs of outright corruption, which is possible only with discretionary centralized authority: The central authority may seek bribes or other favors and may block efficient decisions when bribes are not paid. Or, the authority may favor an inefficient supplier who offers a bribe over a more efficient supplier who does not. Bribery scandals involving public officials are frequently reported, as are cases of sexual harassment with bosses demanding sexual favors from candidates for promotion. Among the legal forms of bribery in the United States are the gifts many companies give to their customers' executives (unless the customer is a government entity). Wherever there is discretionary authority over decisions that people care about, there is a temptation to offer or solicit bribes.

Note that monetary bribes themselves do not necessarily represent an economic inefficiency, because they are but transfers. Rather, the costs of corruption arise first because productive decisions are distorted, either from favoring those who pay bribes or from punishing those who refuse. Secondly, if trust between individuals and faith in the system facilitate economic activity, widespread corruption and bribery may result in further, less direct, but possibly more significant costs.

These costs of discretionary authority depend on flaws in decision makers' incentives, intelligence, or character. Presumably, then, they can be reduced or even eliminated by vesting authority in honest, wise individuals and by giving them incentives to care about organizational performance.³⁷ However, discretionary authority results in a second kind of cost which is incurred even when the central authority is both incorruptible and intelligent enough not to interfere in operations without good reason. These are what we call *influence costs*.

Influence costs arise first because individuals and groups within the organization expend time, effort, and ingenuity in attempting to affect others' decisions to their benefit and secondly because inefficient decisions result either directly from these influence activities or, less directly, from attempts to prevent or control them.

At first blush, it might seem easy to avoid these costs: Simply have decision makers ignore attempts at influence. If this does not provide a sufficient incentive to deter influence activities, severely punish any such behavior. In some circumstances, this may in fact be possible, and

we will assume that organizations follow this policy whenever feasible. However, an essential difficulty exists with such an approach. The policy of ignoring attempts at influence – and, indeed, the policy of selective intervention more generally – is not what macroeconomists call “dynamically consistent” or what game theorists call “subgame perfect.” *Ex post*, when relevant information is available and those at lower levels have already taken actions that cannot be reversed, there will be interventions that are now organizationally desirable and that the center will thus want to take. However, recognition of the center's *ex post* incentives will alter the behavior of the organization's members in ways that are organizationally dysfunctional. Thus, the center would like to be able to commit *ex ante* to not making these interventions – that is, to restrict its own discretion. For example, decision makers might want to motivate workers by committing to promote the most productive one. However, after the fact, they would want to renege and promote the worker who, on the basis of training and other credentials, appears best qualified. As long as central decision makers reserve for themselves the right to make selective interventions, commitments are impossible, if only because of the impossibility of complete contracting. Thus, the possibility of attempts at influence will remain and will inevitably exert costs (Milgrom and Roberts 1988a).

One reason influence is inevitable is that decision makers must rely on others for information that is not easily available to them directly. Central office executives are not islands unto themselves; they commonly rely extensively on others for information, suggestions, and analyses to reach decisions.³⁸ Moreover, the employees affected by a decision are often the very ones executives must rely on. In such circumstances, employees will have strong reasons to try to influence decisions, and their attempts at influence will impose costs on the organization. For example, employees may distort the information they report or withhold information from the central office and from other employees. Candidates for possible promotions may spend valuable time polishing their credentials, thereby establishing their qualifications for the desired assignment at the expense of current performance (Milgrom and Roberts 1988a). Managers, worried about how higher authorities will evaluate their performance, may avoid risky but profitable investments because such investments pose career risks if they turn out badly (Holmstrom 1982). Or, less specifically, employees may simply waste time trying to figure out what issues are on the agenda, how they might be personally affected, and how to shape decisions to their benefit. The loss of productivity from these distortions in the way employees spend their time, report their information, and make their de-

cisions is one category of influence costs. These are costs of discretionary authority because they arise only when an authority exists whose decisions can be influenced.

A second sort of influence cost arises when central authorities make suboptimal decisions because of employees' influence activities, particularly their suppressing or distorting information. In some situations these distortions may be undone by properly accounting for individuals' incentives, and efficient decisions may still be reached (Milgrom and Roberts 1986). However, when these incentives are unclear or when the underlying information is so complex that unscrambling is impossible, decision makers will have to rely on information that they know is incomplete or inaccurate. Consider, for example, the problems of the U.S. Congress in dealing with military appropriations. Congress must rely on the military for information, and it understands that the military may have incentives to distort the information that it provides. But it is impossible for Congress to disentangle interservice rivalries, individual career ambitions, and genuine concerns with national security, all of which motivate particular spending requests. Even if the incentives of those providing the information to distort or suppress it could be determined, the impossible problem of inferring what information they actually have would still remain. In such complex circumstances, decisions must be based on fundamentally incorrect information, and inefficient decisions must be expected.

The incentives to attempt to influence an organization's decisions are, to some extent, endogenous. The costs and benefits of influence activities depend on an organization's information-gathering and decision-making procedures and on its reward systems. Thus, careful organizational design can at least partially control the direct costs of influence activities. For example, Holmstrom and Ricart (1986) have investigated how capital budgeting practices that reward investment and growth per se and establish high internal hurdle rates for investments can help alleviate managers' natural reluctance to undertake risky but profitable investments. Milgrom (1988) and Milgrom and Roberts (1988a) have examined how compensation and promotion policies can be used to make employees more nearly indifferent about company decisions, thereby reducing resistance to change and other organizationally unproductive influence activities.³⁹ As an alternative to using compensation policies and promotion criteria to control incentives to attempt influence, Milgrom and Roberts also explored limiting communication between decision makers and potentially affected parties and otherwise restricting these parties' involvement in decision making.

Even the very boundaries of the firm can become design variables

used to control influence. The widespread practice of spinning off or isolating unprofitable subsidiaries can be partly interpreted in these terms: It is done to prevent the subsidiary's employees and management from imposing large influence costs on the organization through attempts to claim corporate resources to cover their losses and thereby to avoid having to become efficient or to curtail operations.⁴⁰ Similarly, a university's policy of requiring its schools to be "tubs on their own bottoms," each individually responsible for its revenues and expenditures (subject to formula payments to or from the central administration), limits influence activities that would amount to raids on other schools' or the university's resources. For example, when universities centrally determine and fund salaries, research support, and teaching loads, faculties have incentives to try to get more for themselves from the center by invoking comparisons with other schools and departments rather than by raising their own resources. Of course, they can (and do) still make the same complaints when financial boundaries exist between schools, but they have less to gain by doing so because resources cannot easily be shifted from the envied to the envious.

Of course, such responses as these bring costs of their own. Worthwhile investments will be foregone, and managers may seek out the wrong investment opportunities; less qualified people will be assigned to key positions; too many valued employees will quit to increase their pay; bad decisions will be made because communication has been restricted and available information is not used; and desirable resource transfers between divisions will not be effected. These costs of employing policies and organizational structures that would be inefficient if influence activities were not a problem are then in themselves a third category of influence costs.

In this context, an important element of organizational design involves trading off these various costs. For example, Japanese firms make use both of wage policies and of organizational rules to facilitate extensive involvement of their employees in decision making without encouraging excessive attempts at influence. Lifetime employment for key decision makers, relatively small wage differentials within age cohorts, relatively low wages for senior executives,⁴¹ and promotions based largely on seniority⁴² combine to insulate employees from the effects of the firm's investment and promotion decisions and to make promotion decisions relatively immune to influence.⁴³

A central example to test the applicability of these ideas against is the case in which a multidivisional conglomerate buys another firm and resolves to run it as an independent division. For our purposes, a firm is a business organization with a central office that has substantial discretionary authority as well as substantial independence from other

discretionary authorities. Expanding the activities carried out within the firm, rather than through the market, increases the range over which centralized discretionary authority may be exercised and, by our logic, should increase the attendant costs. A true conglomerate acquisition is a particularly clean example because there is a clear increase in centralization free of the confounding effects that come from the acquirer's attempts to integrate the acquired firm's assets and operations with its own.

Such acquisitions often fail (Porter 1987), and frequently the acquired division's performance deteriorates. Tenneco's late-1980 acquisition of Houston Oil and Minerals Corporation is illustrative.⁴⁴ Although Tenneco (then America's largest conglomerate) had resolved to run Houston as an independent subsidiary, within a year of the acquisition Tenneco lost 34 percent of Houston's management, 25 percent of its explorationists, and 19 percent of its production people. All this made it impossible for Tenneco to maintain Houston as a distinct unit. A Tenneco executive commented on the difficulties occasioned by the acquisition of Houston, which was accustomed to giving large production-related bonuses to key people: "We have to ensure internal equity and apply the same standard of compensation to everyone." Why did this acquisition fail? And why did the executive insist on the need for "equity" and a commonly applied "standard of compensation"?

In an acquisition like that of Houston Oil, the acquired firm's previously independent chief executive is replaced by a division head subordinated to the larger organization's central office. This opens up several new kinds of interventions for the conglomerate chief, each of which carries costs of the kind already described. With new levels of executives having authority, there are greater possibilities for mistaken or self-interested interventions. The opportunities for influence costs to arise also expand. The head of an older division may attempt to influence the chief's new decisions by, for example, demanding that the new division purchase supplies from it. One argument might run that although the old division's prices, based on average costs, make its product unattractive to the new division, internal acquisition still serves the overall firm's interest because marginal production costs are low. Similarly, the head of the new division may play politics in an attempt to influence job assignments, pay, and capital budgeting decisions. These are new and costly uses of executive time that were not incurred in the same form⁴⁵ before the firm was acquired. Finally, division heads will expend some resources on defensive influence. For example, the newly acquired division must be prepared to explain why its positions should be filled by promotion from within or

why its salaries and bonuses – high compared to those in other divisions – should not be part of the larger organization's general salary pool.

Taken together, the activities just described could consume a major portion of division heads', and central office personnel's time, diverting them from more productive activities. The boundaries between independent firms reduce the possibilities for influence.⁴⁶ Consequently, those boundaries reduce influence costs.

In the case of Houston Oil, Tenneco's failure to run Houston as an independent subsidiary can most likely be explained by excessive intervention arising from a combination of mistaken perceptions and influence activities. Tenneco's executives may have seen an opportunity to cut wages or benefits for Houston's generously compensated professional work force, disbelieving Houston's protestations that the results would be disastrous. Or, employees in other divisions may have coveted Houston's compensation package, raising the organization's costs of making an exception for Houston. Either way, the mere existence of an executive with discretionary authority to intervene imposed costs that could have been avoided if Houston had remained separate.

The validity of these arguments depends on our characterization of the firm as a centrally controlled organization considerably free from outside intervention. In capitalist economies, several institutions support executives' having much more extensive control over their firms than do courts or government agencies acting from outside. First, property rights tend to limit government interventions more than executive interventions because property rights over the firm generally reside at the executive level or higher. Thus, a court, a governmental regulatory agency, and a firm's central office can all order a plant that is polluting the environment to cease operations until the problem is fixed, but the central office can also replace the plant manager if it finds that to be the most effective way to do the job. Second, executives generally have better and more fluid information systems than courts or government agencies do. Managers in firms hear most of the important information they need in conversations and meetings where they can query sources informally to resolve ambiguities and acquire needed detail.⁴⁷ In contrast, agencies and courts must rely on written reports or adversary proceedings. Finally, executives can deliver incentives directly where they count most – to individual employees – and can tailor the incentives to take the form either of rewards, such as pay increases, bonuses, promotions, or desirable assignments, or of punishments, such as undesirable assignments or layoffs. The incentives courts and government agencies offer consist mostly of threats to collect penalties against the firm's treasury.

Moreover, although some laws explicitly allow discretion to regulators, and others are so vague that the courts have considerable latitude in interpreting them, the role of courts and government agencies is principally to enforce rules. The court or agency must justify its action in terms of the particular rule to be enforced. This procedure denies courts and agencies the degree of fully discretionary authority that a firm's sole proprietor, partners, or senior executives and board can exercise. In fact, this difference of degree is at times so great as to be fairly treated as one of kind.

Still, we do not wish to overstate the extent of centralized authority actually exercised in firms. The most decentralized multidivisional businesses allow division managers considerable autonomy. The holding companies that existed in the United States in the early twentieth century were even more decentralized; their central offices were little more than partial substitutes for capital markets and bankers. However, the authority to intervene, even if not often exercised, still remains and still may exact costs.

Finally, although our argument views firms in a capitalist economy as having considerable autonomy, one should not underestimate the degree of centralized authority present in market economies. Courts do settle contract disputes and interpret the law. Government agencies issue permits, restrain certain business activities, and enforce court orders. Legislatures enact laws to govern contracts, to limit firms' rights to pollute or to engage in dangerous activities, to govern foreign trade, to control the use of land, and to promote societal ends, such as developing the arts or improving the economic status of women and minorities. If our principles are indeed general, then these forms of centralized intervention must be subject to some of the same costs that accompany the creation of centralized executive authority within firms.

INFLUENCE COSTS IN THE PUBLIC SECTOR

Our theory of influence costs dovetails with the theory of rent-seeking behavior. The seminal essays exploring this theory are those by Tullock (1967), Krueger (1974), and Posner (1975), all of which are reprinted in Buchanan, Tollison, and Tullock (1980). The theory holds that government interventions in the economy, whether in the form of tariffs, regulations, the awarding of monopoly franchises, or various attempts to correct market failures, are costly because they create rents and so lead firms and citizens to waste resources attempting to capture those rents. Although this argument has obvious appeal, its presumption that rents lead to inefficiencies only when they result from *government* in-

Bargaining, influence costs, and organization

tervention (as in Buchanan's essay in Buchanan et al. 1980) is, we believe, a mistake. Our general proposition is that *any* centralization of authority, whether in the public or private sector, creates the potential for intervention and so gives rise to costly influence activities and to excessive intervention by the central authority. These costs need to be weighed against the benefits of centralization to determine the efficient extent and locus of authority.⁴⁸

Of course, our theoretical argument that increased centralization leads to increased influence applies with as much force to government and nonprofit organizations as to firms. As an empirical matter, then, we should look for influence activities and their costs in the halls of government as well as in the executive offices of firms. Instances of influence in government are not difficult to find. The frustration of U.S. federal officials who try to manage the nation's affairs in the face of constant attempts at influence was highlighted in recent testimony by former U.S. Secretary of State George Shultz: "Nothing ever gets settled in this town. It's not like running a company or even a university. It's a seething debating society in which the debate never stops, in which people never give up, including me, and that's the atmosphere in which you administer."⁴⁹

The current crisis in tort litigation in the United States provides a second illustration of the importance of influence costs in government. The crisis has arisen in part from the increasing frequency with which novel legal arguments win. In effect, courts increasingly act like discretionary authorities, and litigants incur costs in their efforts to capture newly appropriable sums. The costs of this litigation, which diverts some of the nation's finest minds into largely nonproductive activities and causes talented corporate executives to devote much of their time to defending and avoiding lawsuits, are enormous. The offsetting gains, in improved justice, for example, are much harder to estimate. Limits on damage awards, which are puzzling in standard economic theory,⁵⁰ are easily understood as a device to reduce influence costs.

The importance for encouraging economic development of limiting government's discretionary authority is clear in the economic history of Western Europe. Rosenberg and Birdzell (1986, p. 113) have identified these limits as among the key factors that encouraged the development of trade and early capitalism:

Some of the institutional innovations reduced the risks of trade, either political or commercial. Among them were a legal system designed to give predictable, rather than discretionary, decisions; the introduction of bills of exchange, which facilitated the transfer of money and provided the credit need for commercial transactions; the rise of an insurance market; and the change of gov-

ernmental revenue systems from discretionary expropriation to systematic taxation – a change closely linked to the development of the institution of private property.

Constitutional checks on governmental power that limit both what interventions can be made and who can make them thus reduce the costs of centralized authority. The improved economic efficiency that can accompany constitutional limitations on state power can be spectacular, as in the case of the gains that followed the Glorious Revolution in England (North and Weingast 1987). Similar limitations enforced within private organizations presumably have similar effects. For example, union contracts that govern layoffs and job assignments or antidiscrimination laws may improve efficiency by restricting managerial discretion.

Of course, rules themselves must be decided upon – either centrally or through bargaining; presumably their general applicability renders the stakes large. However, to the extent that rules can be set up well in advance of their application, so that their *predictable* distributional consequences are small, then the bargaining and influence costs incurred in rule making may be small relative to the potential gains. For this to be true, constitutional change must be a difficult and slow process or must require near unanimity among the affected parties.

SUMMARY AND CONCLUSIONS

We have examined the organization of economic activity under the hypothesis that capitalist economic institutions are organized so as to minimize the sum of the costs of resources used in production and the costs of managing the necessary transactions. The costs of negotiating short-term contracts emerged as the distinctive costs of traditional market transactions. An analysis of the determinants of these bargaining costs indicates that two leading theories, one attributing transaction costs primarily to specialized assets and the other attributing them primarily to measurement costs, could both be subsumed under the bargaining cost approach.

The costs associated with nonmarket forms of organization have received less attention in the existing literature, but must be assessed to identify when market organization is more economical than internal procurement. As a first step, we argued that transactions within firms in a capitalist economy are characterized by greater centralization of authority than market-mediated transactions. Indeed, top management's autonomy and discretion and lower management's lesser autonomy are the firm's principal defining characteristics.

Whenever a central authority, whether a governmental unit or an executive in a firm, has discretion to intervene, certain identifiable costs are incurred. These include (1) a tendency for the authority to intervene excessively, both because intervening is that authority's job and because the authority may have a personal interest (licit or illicit, but in either case differing from the organization's interests) in certain decisions; (2) increased time devoted to influence activities and a corresponding reduction in organizational productivity, as interested parties seek to have the authority intervene in particular ways or to adopt their favored alternatives; (3) poorer decision making resulting from the distortion of information associated with influence activities; and (4) a loss of efficiency as the organization adapts its structure and policies to control influence activities and their costs.

We believe that these ideas about influence cost are important in analyzing organizations. For example, they might be used to examine issues of corporate control, financial structure, bankruptcy, proxy fights, and takeovers. Moreover, because influence activities are essentially political and because the theory applies equally to public and private organizations, we believe that it may also prove valuable in the more general study of political economy.