3

Transaction Cost Economics

1. Introduction

Recent and continuing headway notwithstanding, transaction cost economics maintains that our understanding of the economic institutions of capitalism—firms, markets, hybrids, bureaus—is very primitive. It subscribes to the following modest research objective: "to organize our necessarily incomplete perceptions about the economy, to see connections that the untutored eye would miss, to tell plausible . . . causal stories with the help of a few central principles, and to make rough quantitative judgments about the consequences of economic policy and other exogenous events" (Solow, 1985, p. 329).

Transaction cost economics adopts a contractual approach to the study of economic organization. Questions such as the following are germane: Why are there so many forms of organization? What main purpose is served by alternative modes of economic organization and best informs the study of these matters? Striking differences among labor markets, capital markets, intermediate product markets, corporate governance, regulation, and family organization notwithstanding, is it the case that a common theory of contract informs all? What core features—in human, technology, and process respects—does such a common theory of contract rely on? These queries go to the heart of the transaction cost economics research agenda.

The background out of which transaction cost economics works is sketched in Section 2. The operationalization of transaction cost economics is discussed in Section 3. Vertical integration, an understanding of what serves as a paradigm for helping to unpack the puzzles of complex economic organization more generally, is the subject of Section 4. Other applications of the transaction cost approach are examined in Section 5. Some empirical tests of the transaction cost hypotheses are briefly summarized in Section 6. Public policy ramifications are developed in Section 7. Concluding remarks follow.

2. Background

2.1. Main Case

Economic organization services many purposes. Among those that have been ascribed by economists are monopoly and efficient risk bearing. Power and associational gains are sometimes held to be the main purposes of economic organization, especially by noneconomists. And some hold that "social institutions and arrangement . . . [are] the adventitious result of legal, historical, or political forces" (Granovetter, 1985, p. 488).

The study of complex systems is facilitated by distinguishing core purposes from auxiliary purposes. Transaction cost economics subscribes to and develops the view that economizing is the core problem of economic organization.

Main case frameworks do not purport to be exhaustive but are designed to go to the fundamentals.[1] Especially in an area where opinions proliferate, of which the economics of organization is one, insistence upon refutable implications is needed to sort the wheat from the chaff. This is the touchstone function to which Georgescu-Roegan refers (1971, p. 37).

2.2. Behavioral Assumptions

Many economists treat behavioral assumptions as unimportant. This reflects a widely held opinion that the realism of the assumptions is unimportant and that the fruitfulness of a theory turns on its implications (Friedman, 1953). But whereas transaction cost economics is prepared to be judged (comparatively) by the refutable implications which this approach uniquely affords, it also maintains that the behavioral assumptions are important—not least of all because they serve to delimit the study of contract to the feasible subset.

Knight insisted that the study of economic organization needed to be informed by an appreciation for "human nature as we know it" (1965, p. 270), with special reference to the condition of "moral hazard" (1965, p. 260). And Bridgeman reminded social scientists that "the principal problem in understanding the actions of men is to understand how they think—how their minds work" (1955, p. 450). Coase more recently remarked that "modern institutional economics should start with real institutions. Let us also start with man as he is" (1984, p. 231). Coase urges in this connection that the view of man as a "rational utility maximizer" should be abandoned (1984, p. 231), but the salient attributes of "man as he is" otherwise remain undescribed.

I have previously argued that contracting man is distinguished from the orthodox conception of maximizing man in two respects. The first of these is the condition of bounded rationality. Second, contracting man is given to self-

1. Agreement on the main case does not imply that extensions to the main case, to make allowance, for example, for monopoly purposes (where the appropriate preconditions hold), cannot be made. But this is very different from making monopoly the main case—to which economizing is an added wrinkle

interest seeking of a deeper and more troublesome kind than his economic man predecessor.

Although it is sometimes believed that Herbert Simon's notion of bounded rationality is alien to the rationality tradition in economics, Simon actually enlarges rather than reduces the scope for rationality analysis. Thus, the economic actors with whom Simon is concerned are "*intendedly* rational, but only *limitedly* so" (Simon, 1961, p. xxiv). Both parts of the definition warrant respect. An economizing orientation is elicited by the intended rationality part of the definition, while the study of institutions is encouraged by acknowledging that cognitive competence is limited: "It is only because individual human beings are limited in knowledge, foresight, skill, and time that organizations are useful investments for the achievement of human purpose" (Simon, 1957b, p. 199).

Transaction cost economics pairs the assumption of bounded rationality with a self-interest-seeking assumption that makes allowance for guile. Specifically, economic agents are permitted to disclose information in a selective and distorted manner. Calculated efforts to mislead, disguise, obfuscate, and confuse are thus admitted. This self-interest-seeking attribute is variously described as opportunism, moral hazard, and agency.[2]

Bounded rationality and opportunism serve both to refocus attention and help to distinguish between feasible and infeasible modes of contracting. Both impossibly complex and hopelessly naive modes of contracting are properly excluded from the feasible set. Thus:

1. Incomplete contracting. Although it is instructive and a great analytical convenience to assume that agents have the capacity to engage in comprehensive ex ante contracting (with or without private information), the condition of bounded rationality precludes this. All contracts within the feasible set are incomplete. Accordingly, the ex post side of a contract takes on special economic importance. The study of structures that facilitate gapfilling, dispute settlement, adaptation, and the like thus become part of the problem of economic organization. Whereas such institutions play a central role in the transaction cost economics scheme of things, they are ignored (indeed, suppressed) by the fiction of comprehensive ex ante contracting.[3]

2. Critics of transaction cost economics sometimes characterize it as "neo-Hobbesian" because it assumes that economic agents are given to opportunism (in varying degrees). See, for example, Bowles and Gintis (1986, p. 201). Note, however, that the bilateral design of credible commitments (as well as other forms of private ordering) is a very non-Hobbesian response.

3. Note, moreover, that impossibly complex contracting processes cannot be saved by invoking economic natural selection arguments. Natural selection applies only to the set of viable practices and cannot be used to extend the domain. Alchian's claim that "the economist, using the present analytical tools developed in the analysis of the firm under certainty, can predict the more adoptable or viable types of economic interrelationships that will be induced by environmental change even if individuals themselves are unable to ascertain them" (1950, p. 218) is both prescient and provocative. But the argument needs to be invoked with care (Nelson and Winter, 1982). Thus, whereas it is plausible to invoke natural selection to support an efficient factor proportions outcome in a competitively organized industry (Becker, 1962), since the choice of efficient proportions—by accident, insight, or otherwise—by some subset of firms is entirely feasible, to invoke natural selection to support a vaguely described process of "ex post settling

2. Contract as promise. Another convenient concept of contract is to assume that economic agents will reliably fulfill their promises. Such stewardship behavior will not obtain, however, if economic agents are given to opportunism. Ex ante efforts to screen economic agents in terms of reliability and, even more, ex post safeguards to deter opportunism take on different economic significance as soon as the hazards of opportunism are granted. Institutional practices that were hitherto regarded as problematic are thus often seen to perform valued economizing purposes when their transaction cost features are assessed.

Inasmuch as alternative theories of contract with different behavioral assumptions support different definitions of the feasible set, rival theories of contact can, in principle, be evaluated by ascertaining which of the implied feasible sets is borne out in the data.

2.3. Legal Centralism Versus Private Ordering

It is often assumed, sometimes tacitly, that property rights are well defined and that the courts dispense justice costlessly. The mechanism design literature expressly appeals to the efficacy of court ordering (Baiman, 1982, p. 168). Much of the legal literature likewise assumes that the appropriate legal rules are in place and that the courts are the forum to which to present and resolve contract disputes.

The attractions of legal centralism notwithstanding, this orientation was disputed by Llewellyn (1931). He took exception to prevailing contract law doctrine, which emphasized legal rules, and argued that more attention should be given to the purposes served. Less concern with form and more with substance was thus indicated—especially since being legalistic could stand in the way of getting the job done. A rival conception of "contract as framework" was advanced.

If, as Galanter has subsequently argued, the participants to a contract can often "devise more satisfactory solutions to their disputes than can professionals constrained to apply general rules on the basis of limited knowledge of the dispute" (1981, p. 4), then court ordering is better regarded as a background factor rather than the central forum for dispute resolution. Albeit useful for purposes of ultimate appeal, legal centralism (court ordering) gives way to private ordering. This is intimately connected to the incomplete contracting/ex post governance approach to which I refer above.

---

up," whereby managers are purportedly paid their individual marginal products (Fama, 1980), is highly problematic. Unless and until *feasible process mechanics* are described, ex post settling up, at least in its stronger forms, looks like and performs the functions of a deus ex machina.

This is not, however, to say that natural selection plays no role in the study of contract. To the contrary, transaction cost economics maintains that those forms of organization that serve to economize on bounded rationality and safeguard transactions against the hazards of opportunism will be favored and will tend to displace inferior modes in these respects. But transaction cost economics insistently deals only with feasible modes. Within this subset it focuses analytic attention on those properties of organization that have economizing and safeguarding features.

3. Operationalizing Transaction Cost Economics

As elaborated elsewhere (Williamson, 1985b, pp. 2–7), the decade of the 1930s recorded striking insights—in law, economics, and organization—on which transaction cost economics has subsequently built. A thirty-five year interval elapsed, however, during which time the transaction cost approach to economic organization languished and the applied price theory approach to Industrial Organization ruled the day (Coase, 1972, pp. 63–64). The significant accomplishments of the firm-as-production-function approach notwithstanding, orthodox analysis ignored both the internal organization of the firm and the private ordering purposes of contract. As a consequence, "very little [was known] about the cost of conducting transactions on the market or what they depend on; we know next to nothing about the effect on costs of different groupings of activities within firms" (Coase, 1972, p. 64).

Lack of progress with transaction cost economics notwithstanding, the intuition that the leading institutions of economic organization had transaction cost origins was widely shared. As Arrow observed, "market failure is not absolute, it is better to consider a broader category, that of transaction costs, which in general impede and in particular cases completely block the formation of markets" (1969, p. 48). It was not, however, obvious how to operationalize this insight.

3.1. The Technology of Transacting

Adopting Commons' proposal that the transaction be made the basic unit of analysis, attention is focused on economizing efforts that attend the organization of transactions—where a transaction occurs when a good or service is transferred across a technologically separable interface. One stage of activity terminates and another begins. With a well-working interface, as with a well-working machine, these transfers occur smoothly. In mechanical systems we look for frictions: do the gears mesh, are the parts lubricated, is there needless slippage or other loss of energy? The economic counterpart of friction is transaction cost: for that subset of transactions where it is important to elicit cooperation,[4] do the parties to the exchange operate harmoniously, or are there frequent misunderstandings and conflicts that lead to delays, breakdowns, and other malfunctions? Transaction cost analysis entails an examination of the comparative costs of planning, adapting, and monitoring task completion under alternative governance structures.

4. The genius of neoclassical economics is that there are large numbers of transactions where conscious cooperation between traders is not necessary. The invisible hand works well if each party can go its own way—the buyer can secure product easily from alternative sources; the supplier can redeploy his assets without loss of productive value—with little cost to the other. Transaction cost economics is concerned with the frictions that obtain when contractual hazards arise by reason of bilateral dependency, leakage, strategizing, or the like.

Assessing the technology of transacting is facilitated by making the transaction the basic unit of analysis. The central question then becomes: What are the principal dimensions with respect to which transactions differ? Refutable implications are derived from the hypothesis that transactions, which differ in their attributes, are assigned to governance structures, which differ in their costs and competencies, in a discriminating—mainly transaction cost economizing—way.

The principal dimensions on which transaction cost economics presently relies for purposes of describing transactions are (1) the frequency with which they recur, (2) the degree and type of uncertainty to which they are subject, and (3) the condition of asset specificity. Although all are important, many of the refutable implications of transaction cost economics turn critically on this last.

3.1.1. Asset specificity

Asset specificity has reference to the degree to which an asset can be redeployed to alternative uses and by alternative users without sacrifice of productive value. This has a relation to the notion of sunk cost. But the full ramifications of asset specificity become evident only in the context of incomplete contracting and went unrecognized in the pre-transaction cost era (Williamson, 1975, 1979a); Klein, Crawford, and Alchian, 1978).

Interestingly, Marshall recognized that idiosyncratic human capital could sometimes accrue during the course of employment (1948, p. 626). Becker (1962), moreover, made express provision for human capital in his examination of labor market incentive schemes. Marschak expressly took exception with the readiness with which economists accept and employ assumptions of fungibility. As he put it, "There exist almost unique, irreplaceable research workers, teachers, administrations; just as there exist unique choice locations for plants and harbors. The problem of unique or imperfectly standardized goods ... has indeed been neglected in the textbooks" (1968, p. 14). Polanyi's (1962) remarkable discussion of "personal knowledge" further illustrates the importance of idiosyncratic knowledge and working relations.

Transaction cost economics accepts all of the foregoing and moves the argument forward in three respects: (1) asset specificity can take many forms, of which human asset specificity is only one; (2) asset specificity not only elicits complex ex ante incentive responses but, even more important, it gives rise to complex ex post governance structure responses; and (3) the study of economic organization in all of its forms—industrial organization, labor, international trade, economic development, family organization, comparative systems, and even finance—becomes grist for the transaction cost economics mill.

Without purporting to be exhaustive, asset specificity distinctions of six kinds have been made: (1) site specificity, as where successive stations are located in a cheek-by-jowl relation to each other so as to economize on inventory and transportation expenses; (2) physical asset specificity, such as

specialized dies that are required to produce a component; (3) human asset specificity that arises in a learning-by-doing fashion; (4) dedicated assets, which are discrete investments in general purpose plant that are made at the behest of a particular customer; to which (5) brand name capital and (6) temporal specificity have been added. As discussed in Sections 4 and 5, the organizational ramifications of each type of specificity differ. Additional predictive content arises in this way.

3.1.2. Uncertainty

Koopmans described the core problem of the economic organization of society as that of facing and dealing with uncertainty (1957, p. 147). He distinguished between primary and secondary uncertainty in this connection, the distinction being that whereas primary uncertainty is of a state-contingent kind, secondary uncertainty arises "from lack of communication, that is from one decision maker having no way of finding out the concurrent decisions and plans made by others"—which he judges to be "quantitatively at least as important as the primary uncertainty arising from random acts of nature and unpredictable changes in consumer's preferences" (pp. 162–63).

Note, however, that the secondary uncertainty to which Koopmans refers is of a rather innocent or nonstrategic kind. There is a lack of timely communication, but no reference is made to strategic nondisclosure, disguise, or distortion of information. Such strategic features are unavoidably presented, however, when parties are joined in a condition of bilateral dependency. A third class of uncertainty—namely, behavioral (or binary) uncertainty—is thus usefully recognized.[5]

The distinction between *statistical risks* and *idiosyncratic trading hazards* is pertinent in this connection. This is akin to, but nonetheless different from, Knight's (1965) distinction between risk and uncertainty. Hazards are due to the behavioral uncertainties that arise when incomplete contracting and asset specificity are joined. Of special importance to the economics of organization is that the mitigation of hazards can be the source of mutual gain. The language of governance, rather than statistical decision theory, applies.

3.1.3. The fundamental transformation

Economists of all persuasions recognize that the terms upon which an initial bargain will be struck depend on whether noncollusive bids can be elicited from more than one qualified supplier. Monopolistic terms will obtain if there is only a single highly qualified supplier, while competitive terms will result if there are many. Transaction cost economics fully accepts this description of ex ante bidding competition but insists that the study of contracting be extended to include ex post features. Thus, initial bidding merely sets the contracting process in motion. A full assessment requires that both contract

5. The recent paper by Helfat and Teece (1987) examines vertical integration with reference to this condition.

execution and ex post competition at the contract renewal interval come under scrutiny.

Contrary to earlier practice, transaction cost economics holds that a condition of large numbers bidding at the outset does not necessarily imply that a large numbers bidding condition will obtain thereafter. Whether ex post competition is fully effacious or not depends on whether the good or service in question is supported by durable investments in transaction specific human or physical assets. Where no such specialized investments are incurred, the initial winning bidder realizes no advantage over nonwinners. Although it may continue to supply for a long period of time, this is only because, in effect, it is continuously meeting competitive bids from qualified rivals. Rivals cannot be presumed to operate on a parity, however, once substantial investments in transaction specific assets are put in place. Winners in these circumstances enjoy advantages over nonwinners, which is to say that parity at the renewal interval is upset. Accordingly, what was a large numbers bidding condition at the outset is effectively transformed into one of bilateral supply thereafter. The reason why significant reliance investments in durable, transaction specific assets introduce contractual asymmetry between the winning bidder on the one hand and nonwinners on the other is because economic values would be sacrificed if the ongoing supply relation were to be terminated.

Faceless contracting is thereby supplanted by contracting in which the pairwise identity of the parties matters. Not only is the supplier unable to realize equivalent value were the specialized assets to be redeployed to other uses, but the buyer must induce potential suppliers to make similar specialized investments were he to seek least-cost supply from an outsider. The incentives of the parties to work things out rather than terminate are thus apparent. This has pervasive ramifications for the organization of economic activity.

3.2. A Simple Contractual Schema

3.2.1. The general approach

Assume that a good or service can be supplied by either of two alternative technologies. One is a general purpose technology, the other a special purpose technology. The special purpose technology requires greater investment in transaction-specific durable assets and is more efficient for servicing steady-state demands.

Using k as a measure of transaction-specific assets, transactions that use the general purpose technology are ones for which $k = 0$. When transactions use the special purpose technology, by contrast, a $k > 0$ condition exists. Assets here are specialized to the particular needs of the parties. Productive values would therefore be sacrificed if transactions of this kind were to be prematurely terminated. The bilateral monopoly condition described above and elaborated below applies to such transactions.

Whereas classical market contracting—"sharp in by clear agreement; sharp out by clear performance" (Macneil, 1974, p. 738)—suffices for transactions of the $k = 0$ kind, unassisted market governance poses hazards when-

ever nontrivial transaction-specific assets are placed at risk. Parties have an incentive to devise safeguards to protect investments in transactions of the latter kind. Let s denote the magnitude of any such safeguards. An $s = 0$ condition is one in which no safeguards are provided; a decision to provide safeguards is reflected by an $s > 0$ result.

Figure 3.1 displays the three contracting outcomes corresponding to such a description. Associated with each node is a price. So as to facilitate comparisons between nodes, assume that suppliers (1) are risk neutral, (2) are prepared to supply under either technology, and (3) will accept any safeguard condition whatsoever so long as an expected breakeven result can be projected. Thus, node A is the general purpose technology ($k = 0$) supply relation for which a breakeven price of p_1 is projected. The node B contract is supported by transaction-specific assets ($k > 0$) for which no safeguard is offered ($s = 0$). The expected breakeven price here is $\bar{p}$. The node C contract also employs the special purpose technology. But since the buyer at this node provides the supplier with a safeguard, ($s > 0$), the breakeven price, $\hat{p}$, at node C is less than $\bar{p}$.[6]

The protective safeguards to which I refer normally take on one or more of three forms. The first is to realign incentives, which commonly involves some type of severance payment or penalty for premature termination. Albeit important and the central focus of much of the formal contracting literature, this is a very limited response. A second is to supplant court ordering by private ordering. Allowance is expressly made for contractual incompleteness; and a different forum for dispute resolution (of which arbitration is an example) is commonly provided (see Joskow, 1985, 1987; Williamson, 1985b, pp. 164–66). Third, the transaction may be embedded in a more complex trading network. The object here is to better assure continuity purposes and facilitate adaptations. Expanding a trading relation from unilateral to bilateral exchange—through the concerted use, for example, of reciprocity—thereby to effect an equilibration of trading hazards is one illustration. Recourse to collective decision-making under some form of combined ownership is another.

This simple contracting schema applies to a wide variety of contracting issues. It facilitates comparative institutional analysis by emphasizing that technology (k), contractual governance/safeguards (s) and price (p) are fully interactive and are determined simultaneously. It is furthermore gratifying that so many applications turn out to be variations on a theme.

By way of summary, the nodes A, B, and C in the contractual schema set out in Figure 3.1 have the following properties:

1. Transactions that are efficiently supported by general purpose assets ($k = 0$) are located at node A and do not need protective governance structures. Discrete market contracting suffices. The world of competition obtains.

6. Specialized production technologies commonly afford steady-state cost savings over general purpose production technologies. But since the former are less redeployable than the latter, stochastic disturbances may reverse the cost advantage (whether p_1 is greater than or less than $\hat{p}$ requires that stochastic factors be taken into account). See Williamson, 1985b, pp. 169–75.

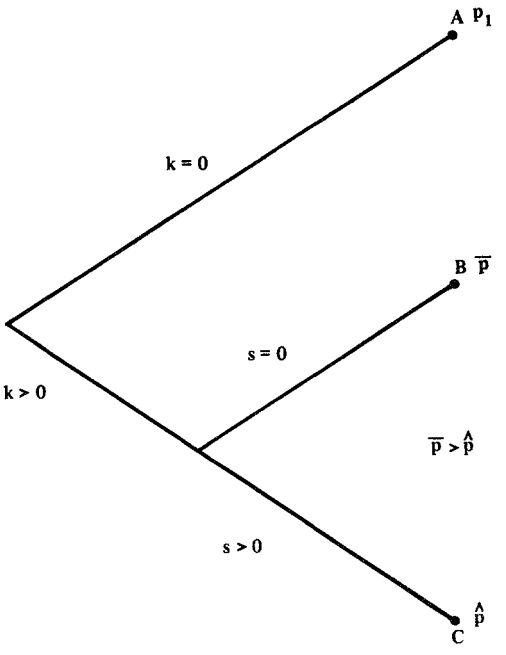


Figure 3.1. A simple contracting schema.

2. Transactions that involve significant investments of a transaction-specific kind ($k > 0$) are ones for which the parties are effectively engaged in bilateral trade.

3. Transactions located at node B enjoy no safeguards ($s = 0$), on which account the projected breakeven supply price is great ($\bar{p} > \hat{p}$). Such transactions are apt to be unstable contractually. They may revert to node A [in which event the special purpose technology would be replaced by the general purpose ($k = 0$) technology] or be relocated to node C (by introducing contractual safeguards that would encourage use of the $k > 0$ technology).

4. Transactions located at node C incorporate safeguards ($s > 0$) and thus are protected against expropriation hazards.

5. Inasmuch as price and governance are linked, parties to a contract should not expect to have their cake (low price) and eat it too (no safeguard). More generally, it is important to study *contracting in its entirety.* Both the ex ante terms and the manner in which contracts are thereafter executed vary with the investment characteristics and the associated governance structures within which transactions are embedded.

3.2.2. *An illustration*

Klein and Leffler (1981) argue that franchisees may be required to make investments in transaction-specific capital as a way by which to safeguard the franchise system against quality shading. As Klein puts it, franchisers can better

> assure quality by requiring franchisee investments in specific . . . assets that upon termination imply a capital loss penalty larger than can be obtained by the franchisee if he cheats. For example, the franchiser may require franchisees to rent from them short term (rather than own) the land upon which their outlet is located. This lease arrangement creates a situation where termination can require the franchisee to move and thereby impose a capital loss on him up to the amount of his initial nonsalvageable investment. Hence a form of collateral to deter franchisee cheating is created. (1980, p. 359)

The arrangement is tantamount to the creation of hostages to restore integrity to an exchange.

That logic notwithstanding, the use of hostages to deter franchisees from exploiting the demand externalities that inhere in brand name capital is often regarded as an imposed (top down) solution. Franchisees are "powerless"; they accept hostage terms because no others are available. Such power arguments are often based on ex post reasoning. That the use of hostages to support exchange can be and often is an efficient systems solution, hence is independent of who originates the proposal, can be seen from the following revised sequence.

Suppose that an entrepreneur develops a distinctive, patentable idea that he sells outright to a variety of independent, geographically dispersed suppliers, each of which is assigned an exclusive territory. Each supplier expects to sell only to the population within its territory, but all find to their surprise (and initially to their delight) that sales are also made to a mobile population. Purchases by the mobile population are based not on the reputation of individual franchisees but on customers' perceptions of the reputation of the system. A demand externality arises in this way.

Thus, were sales made only to the local population, each supplier would fully appropriate the benefits of its promotional and quality enhancement efforts. Population mobility upsets this: because the cost savings that result from local quality debasement accrue to the local operator while the adverse demand effects are diffused throughout the system, suppliers now have an incentive to free ride off of the reputation of the system. Having sold the exclusive territory rights outright, the entrepreneur who originated the program is indifferent to these unanticipated demand developments. It thus remains for the collection of independent franchisees to devise a correction themselves, lest the value of the system deteriorate to their individual and collective disadvantage.

The franchisees, under the revised scenario, thus create an agent to police quality or otherwise devise penalties that deter quality deterioration. One possibility is to return to the entrepreneur and hire him to provide such services. Serving now as the agent of the franchisees, the entrepreneur may undertake a program of quality checks (certain purchasing restraints are introduced, whereby franchisees are required to buy only from qualified suppliers; periodic inspections are performed). The incentive to exploit demand externalities may further be discouraged by requiring each franchisee to post a hostage and by making franchises terminable.

This indirect scenario serves to demonstrate that it is the *system* that benefits from the control of externalities. But this merely confirms that the normal scenario in which the franchiser controls the contractual terms is not an arbitrary exercise of power. Indeed, if franchisees recognize that the demand externality exists from the outset, if the franchiser refuses to make provision for the externality in the original contract, and if it is very costly to reform the franchise system once initial contracts are set, franchisees will bid less for the right to a territory than they otherwise would. It should not therefore be concluded that perceptive franchisers, who recognize the demand externality in advance and make provision for it, are imposing objectionable ex ante terms on unwilling franchisees. They are merely taking steps to realize the full value of the franchise. Here, as elsewhere, contracts must be examined in their entirety.

3.3. The Measurement Branch

Most of the foregoing and most of this chapter deal with the governance issues that arise in conjunction with asset specificity. There is, however, another branch that focuses on problems of measurement. The treatment of team organization by Alchian and Demsetz (1972) in the context of technological nonseparabilities is one example. Barzel's (1982) concerns with product quality is another.

All measurement problems are traceable to a condition of information impactedness—which is to say that either (1) information is asymmetrically distributed between buyer and seller and can be equalized only at great cost or (2) it is costly to apprise an arbiter of the true information condition should a dispute arise between opportunistic parties who have identical knowledge of the underlying circumstances (Williamson, 1975, pp. 31–37). Interestingly, measurement problems with different origins give rise to different organizational responses. Thus, whereas team organization problems give rise to supervision, the classical agency problem elicits an incentive alignment response. Reputation effect mechanisms are responses to quality uncertainty, and common ownership is often the device by which concerns over asset dissipation are mitigated. Plainly, an integrated treatment of governance and measurement is ultimately needed.[7]

4. The Paradigm Problem: Vertical Integration

The leading studies of firm and market organization—in 1937 and over the next thirty-five years—typically held that the "natural" or efficient boundaries

7. Alchian joins the two as follows: "One might . . . define the firm in terms of two features: the detectability of *input* performance *and* the expropriability of quasi-rents of [transaction specific] resources" (1984, p. 39). See also Milgrom and Roberts (1992).

of the firm were defined by technology and could be taken as given. Boundary extension was thus thought to have monopoly origins.[8]

Coase (1937) took exception with this view in his classic article on "The Nature of the Firm." He not only posed the fundamental question: When do firms choose to procure in the market and when do they produce to their own requirements?, but he argued that comparative transaction cost differences explain the result. Wherein, however, do these transaction cost differences reside?

The proposition that asset specificity had significant implications for vertical integration was first advanced in 1971. A comparative institutional orientation was employed to assess when and for what reasons market procurement gives way to internal organization. Given the impossibility of comprehensive contracting (by reason of bounded rationality) and the need to adapt a supply relation through time (in response to disturbances), the main comparative institutional alternatives to be evaluated were between incomplete short-term contracts and vertical integration. Problems with short-term contracts were projected "if either (1) efficient supply requires investment in special-purpose, long-life equipment, or (2) the winner of the original contract acquires a cost advantage, say by reason of 'first mover' advantages (such as unique location or learning, including the acquisition of undisclosed or proprietary technical and managerial procedures and task-specific labor skills)" (Williamson, 1971b, p. 116).

4.1. A Heuristic Model

The main differences between market and internal organization are these: (1) markets promote high-powered incentives and restrain bureaucratic distortions more effectively than internal organization; (2) markets can sometimes aggregate demands to advantage, thereby to realize economies of scale and scope; and (3) internal organization has access to distinctive governance instruments.

Consider the decision of a firm to make or buy a particular good or service. Suppose that it is a component that is to be joined to the mainframe and assume that it is used in fixed proportion. Assume, furthermore, that economies of scale and scope are negligible. Accordingly, the critical factors that are determinative in the decision to make or buy are production cost control and the ease of effecting intertemporal adaptations.

Although the high-powered incentives of markets favor tighter production cost control, they impede the ease of adaptation as the bilateral dependency of the relation between the parties builds up. The latter effect is a consequence of the fundamental transformation that occurs as a condition of asset specificity

8. The main monopoly emphasis was on the use of boundary extension to exercise economic muscle (Stigler, 1951, 1955; Bain, 1968). McKenzie (1951) and others have noted, however, that vertical integration may also be used to correct against monopoly-induced factor distortions. Arguments of both kinds work out of the firm-as-production-function tradition. For a much more complete treatment of vertical integration, see Martin Perry, 1989.

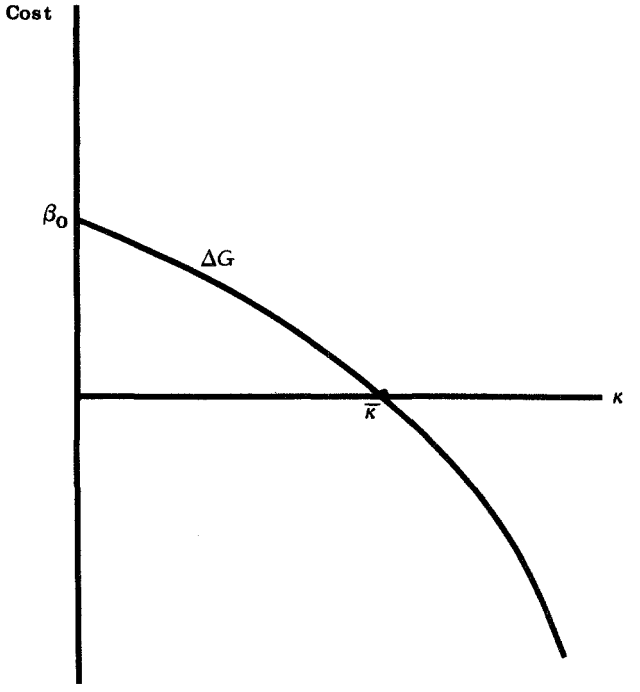


Figure 3.2. Comparative governance cost.

deepens. For a fixed level of output (say $X = \bar{X}$), let $B(k)$ be the bureaucratic costs of internal governance and $M(k)$ the corresponding governance costs of markets, where k is an index of asset specificity. Assume that $B(0) > M(0)$, by reason of the above-described incentive and bureaucratic effects. Assume further, however, that $M' > B'$ evaluated at every k. This second condition is a consequence of the comparative disability of markets in adaptability respects. Letting $\Delta G = B(k) - M(k)$, the relation shown in Figure 3.2 obtains.

Thus, market procurement is the preferred supply mode where asset specificity is slight—because $\Delta G > 0$ under these circumstances. But internal organization is favored where asset specificity is great, because the high-powered incentives of markets impair the comparative ease with which adaptive, sequential adjustments to disturbances are accomplished. As shown, the switchover value, where the choice between firm and market is a matter of indifference, occurs at $\bar{k}$.

The foregoing assumes that economies of scale and scope are negligible, so that the choice between firm and market rests entirely on the governance cost differences. Plainly that oversimplifies. Markets are often able to aggregate diverse demands, thereby to realize economies of scale and scope. Accordingly, production cost differences also need to be taken into account.[9]

9. The argument assumes that the firm produces exclusively to its own needs. If diseconomies of scale or scope are large, therefore, technological features will deter all but very large firms from supplying to their own needs.

Again it will be convenient to hold output unchanged. Let ΔC be the steady-state production cost difference between producing to one's own requirements and the steady-state cost of procuring the same item in the market. (The steady-state device avoids the need for adaptation.) Expressing ΔC as a function of asset specificity, it is plausible to assume that ΔC will be positive throughout but will be a decreasing function of k.

The production cost penalty of using internal organization is large for standardized transactions for which market aggregation economies are great, whence ΔC is large where k is low. The cost disadvantage decreases but remains positive for intermediate degrees of asset specificity. Thus, although dissimilarities among orders begin to appear, outside suppliers are nevertheless able to aggregate the diverse demands of many buyers and produce at lower costs than can a firm that produces to its own needs. As goods and services become very close to unique (k is high), however, aggregation economies of outside supply can no longer be realized, whence ΔC asymptotically approaches zero. Contracting out affords neither scale nor scope economies in those circumstances. The firm can produce without penalty to its own needs.

This ΔC relation is shown in Figure 3.3. The object, of course, is not to minimize ΔC or ΔG taken separately but, given the optimal or specified level of asset specificity, to minimize the sum of production and governance cost differences. The vertical sum $\Delta G + \Delta C$ is also displayed. The crossover value of k for which the sum $(\Delta G + \Delta C)$ becomes negative is shown by $\hat{k}$, which value exceeds $\bar{k}$. Economies of scale and scope thus favor market organization over a wider range of asset specificity values than would be observed if steady state production cost economies were absent.

More generally, if k^* is the optimal degree of asset specificity,[10] Figure 3.3 discloses:

1. Market procurement has advantages in both scale economy and governance respects where optimal asset specificity is slight ($k^* \ll \hat{k}$).

Plausible though this appears, neither economies of scale nor scope are, by themselves, responsible for decisions to buy rather than make. Thus, suppose that economies of scale are large in relation to a firm's own needs. Absent prospective contracting problems, the firm could construct a plant of size sufficient to exhaust economies of scale and sell excess product to rivals and other interested buyers. Or suppose that economies of scope are realized by selling the final good in conjunction with a variety of related items. The firm could integrate forward into marketing and offer to sell its product together with related items on a parity basis—rival and complementary items being displayed, sold, and serviced without reference to strategic purposes.

That other firms, especially rivals, would be willing to proceed on this basis, is surely doubtful. Rather than submit to the strategic hazards, some will decline to participate in such a scheme (Williamson, 1975, pp. 16–19; 1979a, pp. 979–80). The upshot is that *all* cost differences between internal and market procurement ultimately rest on transaction cost considerations. Inasmuch, however, as the needs of empirical research on economic organization are better served by making the assumption that firms which procure internally supply exclusively to their own needs, whence technological economies of scale and scope are accorded independent importance, I employ this assumption here.

10. Reference to a single "optimal" level of k is an expository convenience: the optimal level actually varies with organization form. This is further developed in Subsection 4.2.

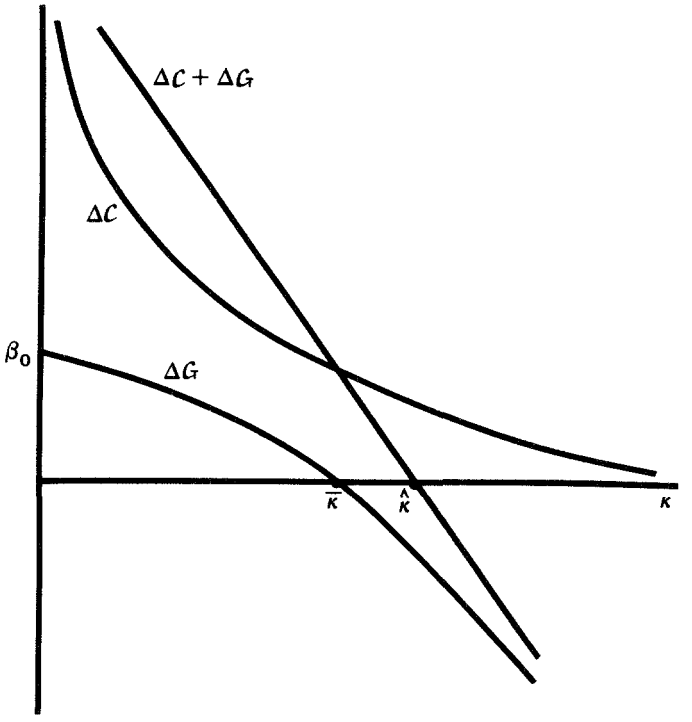


Figure 3.3. Comparative production and governance costs.

2. Internal organization enjoys the advantage where optimal asset specificity is substantial ($k^* \gg \hat{k}$). Not only does the market realize little aggregation economy benefits, but market governance, because of maladaptation problems that arise when assets are highly specific, is hazardous.

3. Only small cost differences appear for intermediate degrees of optimal asset specificity. Mixed governance, in which some firms will be observed to buy, others to make, and all express "dissatisfaction" with their procurement solution, are apt to arise for k^* in the neighborhood of $\hat{k}$. Accidents of history may be determinative. Nonstandard contracts of the types discussed briefly above and developed more fully in Subsection 4.2 may arise to serve these.

4. More generally, it is noteworthy that, inasmuch as the firm is everywhere at a disadvantage to the market in production cost respects ($\Delta C < 0$ everywhere), the firm will never integrate for production cost reasons alone. Only when contracting difficulties intrude does the firm and market comparison support vertical integration—and then only for values of k^* that significantly exceed $\hat{k}$.

Additional implications may be gleaned by introducing quantity (or firm size) and organization form effects. Thus, consider firm size (output). The basic proposition here is that diseconomies associated with own production will be everywhere reduced as the quantity of the component to be supplied increases. The firm is simply better able to realize economies of scale as its

own requirements become larger in relation to the size of the market. The curve ΔC thus everywhere falls as quantity increases. The question then is: What happens to the curve ΔG? If this twists about $\bar{k}$, which is a plausible construction,[11] then the vertical sum $\Delta G + \Delta C$ will intersect the axis at a value of k that progressively moves to the left as the quantity to be supplied increases. Accordingly:

5. Larger firms will be more integrated into components than will smaller, ceteris paribus.

Finally, for reasons that have been developed elsewhere (Williamson, 1970), the bureaucratic disabilities to which internal organization is subject vary with the internal structure of the firm. Multidivisionalization, assuming that the M-form is feasible, serves as a check against the bureaucratic distortions that appear in the unitary form (U-form) of enterprise. Expressed in terms of Figure 3.3, the curve ΔG falls under multidivisionalization as compared with the unitary form organization. Thus, assuming ΔC is unchanged:

6. An M-form firm will be more integrated than its U-form counterpart, ceteris paribus.

4.2. A Combined Neoclassical–Transaction Cost Treatment

A unified framework is herein employed to formalize the arguments advanced above.[12] It is in the spirit of Arrow's remark that new theories of economic organization takes on greater "analytic usefulness when these are founded on more directly neoclassical lines" (1985b, p. 303). The spirit of the analysis is consonant with that of economics quite generally: use more general modes of analysis as a check on the limitations that inform more specialized types of reasoning.

The heuristic model assumes that both firm and market modes of supply produce the same level of output and that the optimal level of asset specificity is the same in each. These are arbitrary constraints, however. What happens when both are relaxed? This is examined below in the context of a combined production and transaction cost model that is itself highly simplified—in that it (1) deals only with polar firm or market alternatives, (2) examines only one transaction at a time, and (3) employs a reduced form type of analysis, in that it ascribes rather than derives the basic production and governance cost competencies of firms and markets. (See, however, Chapter 4.)

It will facilitate the argument to assume initially that firm and market employ the identical production cost technology. This assumption is subsequently relaxed.

11. Assume that $I(k, X) = I(k)X$ where $I(0) > 0$ and $I(k)$ is the internal governance cost per unit of effecting adaptations. Assume, furthermore, that $M(k, X) = M(k)X$ where $M(0) = 0$ and $M(k)$ is the corresponding governance cost per unit of effecting market adaptations. Then $\Delta G = [I(k) - M(k)]X$, and the value at which ΔG goes to zero will be independent of X. The effect of increasing X is to twist ΔG clockwise about the value of k at which it goes to zero.

12. The argument is based on Riordan and Williamson, 1985. See also Masten, 1982.

4.2.1. Common production technology

Revenue is given by $R = R(X)$, and production costs of market and internal procurement are assumed to be given by the relation:

$$C = C(X, k; \alpha); \quad C_X > 0; C_k < 0; C_{Xk} < 0,$$

where the parameter α is a shift parameter, a higher value of α yielding greater cost reducing consequences to asset specificity:

$$C_{k\alpha} < 0; \quad C_{X\alpha} < 0.$$

Asset specificity is assumed to be available at the constant per unit cost of γ. The neoclassical profit expression corresponding to this statement of revenue and production costs is given by

$$\pi^*(X, k; \alpha) = R(X) - C(X, k; \alpha) - \gamma k.$$

Governance costs are conspicuously omitted from this profit relation, there being no provision for such costs in the neoclassical statement of the problem.

Assume that this function is globally concave. At an interior maximum the decision variables X^* and k^* are determined from the zero marginal profit conditions:

$$\pi^*_X(X, k; \alpha) = 0; \quad \pi^*_k(X, k; \alpha) = 0.$$

Consider now the governance costs of internal and market organization. Let the superscripts i denote internal and m denote market organization. Governance cost expressions congruent with the cost differences described above are given by

$$G^{\mathrm{i}} = \beta + V(k); \quad \beta > 0; V_k >= 0,$$
$$G^{\mathrm{m}} = W(k); \quad W_k > 0,$$

where $W_k > V_k$, evaluated at common k.

The corresponding profit expressions for internal market procurement in the face of positive governance costs are

$$\pi^{\mathrm{i}} = R(X) - C(X, k; \alpha) - \gamma k - (\beta + V(k)),$$
$$\pi^{\mathrm{m}} = R(X) - C(X, k; \alpha) - \gamma k - W(k).$$

The zero marginal profit conditions for internal procurement are

$$\pi^{i}_{X} = R_X - C_X = 0,$$
$$\pi^{i}_{k} = -C_k - \gamma - V_k = 0.$$

Those for market procurement are

$$\pi^{m}_{k} = R_X - C_X = 0,$$
$$\pi^{m}_{k} = -C_k - \gamma - W_k = 0.$$

In each instance, therefore, optimal output, given asset specificity, is obtained by setting marginal revenue equal to the marginal costs of production, while optimal asset specificity, given output, is chosen to minimize the sum of production and governance costs.

Given that $\pi^{*}_{Xk} = -C_{Xk} \geq 0$, the neoclassical locus of optimal output given asset specificity and the corresponding locus of optimal asset specificity given output will bear the relations shown by $\pi^{*}_{X} = 0$ and $\pi^{*}_{k} = 0$ in Figure 3.4. The corresponding loci for internal and market organization are also shown. Inasmuch as the zero marginal profit expressions for output for all three statements of the maximand are identical, the loci $\pi^{i}_{X} = 0$ and $\pi^{m}_{X} = 0$ track $\pi^{*}_{X} = 0$ exactly. The zero marginal profit expressions for asset specificity, however, differ. Given that $W_k > V_k > 0$, the locus $\pi^{m}_{k} = 0$ is everywhere below $\pi^{i}_{k} =$

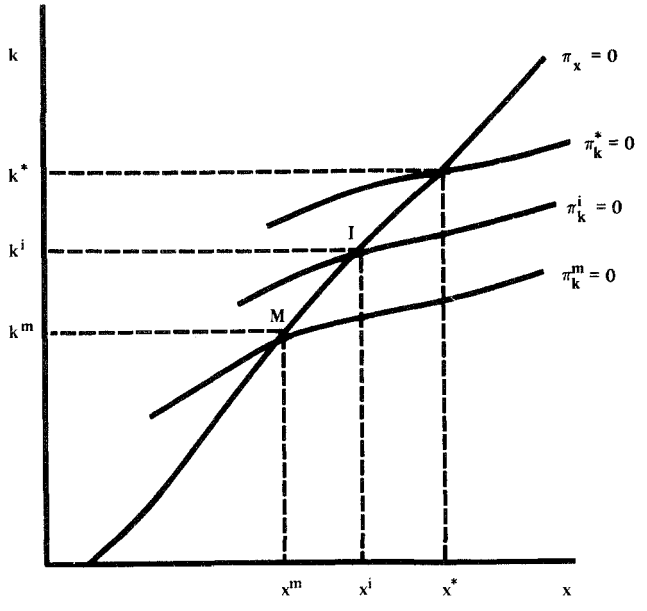


Figure 3.4. Marginal profit loci.

0, which in turn is below $\pi_k^* = 0$. Accordingly, profit maximizing values of X and k for these three statements of the optimization problem bear the following relation to each other: $X^* > X^i > X^m$ and $k^* > k^i > k^m$. The output effects are indirect or induced effects, attributable to shifts in the zero marginal profit asset specificity loci.

Of course, the X^* and k^* choices are purely hypothetical since, in reality, a zero transaction cost condition is not a member of the feasible set. The relevant choices thus reduce to using input combinations I under internal procurement or M under market procurement. An immediate implication is that if the firm were operating in two identical markets and was constrained to buy in one and to make in the other, it would sell more goods of a more distinctive kind in the region where it produced to its own needs.

Ordinarily, however, the firm will not be so constrained but will choose to make or buy according to which mode offers the greatest profit in each region. Figure 3.5 shows profit as a function of asset specificity, the choice of output assumed to be optimal for each value of k. Whereas there is a family of π^i curves, one for each value of the bureaucratic cost parameter β, there in only a single π^m curve. Which mode is favored depends on which has the highest peak. This is the internal mode for $\beta = \beta_0$ but the market mode for $\beta = \beta_1$, where $\beta_1 > \beta_0$. The optimal values of k and X depend only on the mode selected and not on β, however, since β does not influence the marginal conditions.

The comparative statics ramifications of the production cost parameter α are more central. Applications of the envelope theorem reveal that

$$\pi_\alpha^m = -C_\alpha(X^m, k^m; \alpha),$$

$$\pi_\alpha^i = -C_\alpha(X^i, k^i; \alpha).$$

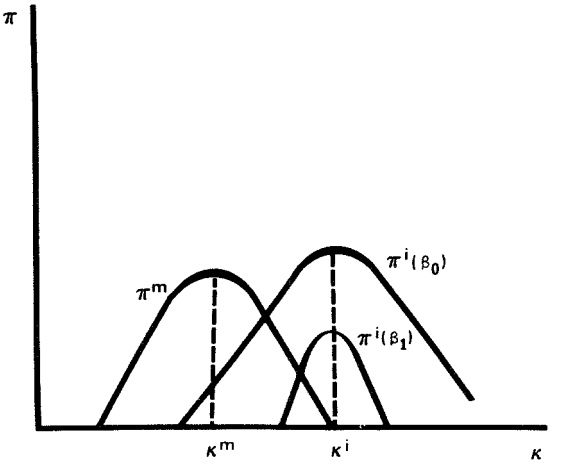


Figure 3.5. Bureaucratic cost effects.

Inasmuch as $X^i > X^m$ and $k^i > k^m$, it follows from our earlier production cost assumptions that $\pi^i_\alpha > \pi^m_\alpha$. In other words, as asset specificity has greater cost reducing impact, internal organization is progressively favored.

4.2.2. *Production cost differences*

Consider now the case, to which earlier reference was made and is arguably the more realistic, where the firm is unable to aggregate demands and sell product that exceeds its own demands without penalty. Let $H(X,k)$ denote the production cost disadvantage per unit of output associated with internal organization. The production costs of the two modes then are

$$C^m = C(X, k; \alpha),$$
$$C^i = C(X, k; \alpha) + H(X, k)X.$$

Assume that $H_X < 0$ and $H_k < 0$ but that $H(X, k)X$ is positive and asymptotically approaches zero as X and k approach infinity. Denote the marginal production cost disadvantage by $M(X, k) = H_X(X, k)X + H(X, k)$.

The analysis depends on the way in which the total production cost disadvantage experienced by internal organization changes for outputs within the relevant range. At low levels of output, decreasing unit cost disadvantages will normally be attended by an increasing total cost, whence $M(X, k) > 0$. Beyond some threshold level of output, however, the total production cost disadvantage of internal organization will begin to decline. Indeed, as the firm progressively increases in relation to the size of the market, the production cost disadvantage presumably approaches zero—since firm and market have access to identical economies of scale as a monopoly condition evolves. Accordingly, $M(X, k) < 0$ once this threshold is crossed.

The main results are strengthened within the (large output) range where $M(X, k) < 0$: $X^m < X^i$, $k^m < k^i$; and $\pi^i_a < \pi^m_a$. Within the (small output) range, however, where $M_x > 0$, the marginal production cost disadvantage of internal organization and the marginal governance cost disadvantage of market procurement operate in opposite directions. An unambiguous ordering of optimal output and asset specificity is not possible in terms of the above-described qualitative features of the problem in this instance. An anomaly thus arises that was not evident in the heuristic presentation above.

5. Other Applications

The underlying transaction cost economizing theme repeats itself, with variation, almost endlessly. Three applications are sketched here: to nonstandard commercial contracting, career marriages, and corporate finance.[13] The sys-

13. Applications to labor market organization and comparative economic systems are developed in Williamson, 1985b, chaps. 9 and 10.

tems' ramifications of organizational innovation are also noteworthy. These are examined with reference to "Full Functionalism" (Elster, 1983).

5.1. Nonstandard Commercial Contracting

Many nonstandard contracting phenomena are explained with the aid of one of two models: the hostage model and the oversearching model.

5.1.1. The hostage model

The hostage model developed in Chapter 5 is a member of the family of models dealing with credible commitments (Klein and Leffler, 1981; Telser, 1981; Williamson, 1983). Although the particulars differ, all of these models feature intertemporal contracting, uncertainty, and investments in transaction specific assets. The application to reciprocal trading is sketched here.

Reciprocity is believed to be a troublesome practice. Reciprocity transforms a unilateral supply relation—whereby A sells X to B—into a bilateral one, whereby A agrees to buy Y from B as a condition for making the sale of X and both parties understand that the transaction will be continued only if reciprocity is observed. Although reciprocal selling is widely held to be anticompetitive (Stocking and Mueller, 1957; Blake, 1973), others regard it more favorably. Stigler offers the following affirmative rationale for reciprocity.

> The case for reciprocity arises when prices cannot be freely varied to meet supply and demand conditions. Suppose that a firm is dealing with a colluding industry which is fixing prices. A firm in this collusive industry would be willing to sell at less that the cartel price if it can escape detection. Its price can be reduced in effect by buying from the customer-seller at an inflated price. Here reciprocity restores flexibility of prices.[14]

Inasmuch, however, as many industries do not satisfy the prerequisites for oligopolistic price collusion (Posner, 1969b; Williamson, 1975, chap. 12) and as reciprocity is sometimes observed among these, reciprocity presumably has other origins as well. Tie breaking is one of these. A second is that reciprocity can have advantageous governance structure benefits. These two can be distinguished by the type of product being sold.

The tie-breaker explanation applies where firm B, which is buying specialized product from A, asks that A buy standardized product from B on the condition that B meets market terms. Other things being equal, procurement agents at A are apt to accede. Scherer notes that "Most of the 163 corporation executives responding to a 1963 survey state that their firms' purchases were awarded on the basis of reciprocity only when the price, quality, and delivery conditions were equal" (1980, p. 344).

14. President's Task Force Report on Productivity and Competition, reprinted in Commerce Clearing House *Trade Regulation Reporter,* June 24, 1969, p. 39.

The more interesting case is where reciprocity involves the sale of specialized product to *B* conditioned on the procurement of specialized product from *B*. The argument here is that reciprocity can serve to equalize the exposure of the parties, thereby reducing the incentive of the buyer to defect from the exchange—leaving the supplier to redeploy specialized assets at greatly reduced alternative value. Absent a hostage (or other assurance that the buyer will not defect), the sale by *A* of specialized product to *B* may never materialize. The buyer's commitment to the exchange is more assuredly signaled by his willingness to accept reciprocal exposure of specialized assets. Defection hazards are thereby mitigated.

Lest the argument be uncritically considered to be a defense for reciprocal trading quite generally, note that it applies only where specialized assets are placed at hazard by both parties. Where only one or neither invests in specialized assets, the practice of reciprocity plainly has other origins.

Shepard (1986) has recently developed another interesting application of transaction cost reasoning that involves not the creation but the release of a hostage. The puzzle to be explained is the insistence by buyers that semiconductor producers license their design of chips to others. One explanation is that this averts delivery failures attributable to idiosyncratic disruptive events at the parent company (earthquakes, labor strife, and the like). If, however, exposure to geographic hazards and supply interruptions due to company-wide bargaining were the only concerns, then subcontracting would afford adequate relief. Since the parent company could retain full control over total production via subcontracting, and since such control offers the prospect of added monopoly gains, licensing is evidently a poorly calibrated—indeed, in relation to the above described economic purposes, it is an excessive—response.

The possibility that the demand for licensing has other origins is thus suggested. The transaction cost rationale for insistence upon licensing is that buyers are reluctant to specialize their product and production to a particular chip without assurance of "competitive" supply. The concern is that a monopoly seller will expropriate the buyer when follow-on orders are placed—which is after the buyer has made durable investments that cannot be redeployed without sacrifice of productive value. The insistence on licensing is thus explained by the fact that access to several *independent* sources of supply relieves these expropriation hazards.[15]

5.1.2. Oversearching

Most of the applications of transaction cost economics have dealt with governance issues. Transaction cost economics also deals, however, with measurement problems (Barzel, 1982). One manifestation of this is oversearching.

15. This is akin to, though slightly different from, Shepard's (1986) explanation.

Kenney and Klein (1983) address themselves to several such cases. One is a reinterpretation of the *Loew's* case,[16] where Kenney and Klein take exception to Stigler's interpretation of block-booking as an effort to effect price discrimination. They argue instead that block-booking economizes on measurement costs for motion picture films the box-office receipts of which are difficult to estimate ex ante.

A more interesting case is their interpretation of the market for gem-quality uncut diamonds. Despite classification into more than two thousand categories, significant quality variation in the stones evidently remains. How can such a market be organized so that oversearching expenses are not incurred and each party to the transaction has confidence in the other? The "solution" that the market evolved and which Kenney and Klein interpret entailed the assembly of groups of diamonds—or "sights"—and imposing all-or-none and in-or-out trading rules. Thus, buyers who refuse to accept a sight are thereafter denied access to this market.

These two trading rules may appear to "disadvantage" buyers. Viewed in systems terms, however, they put a severe burden on de Beers to respect the legitimate expectations of buyers. Thus, suppose that only an all-or-none trading rule were to be imposed. Although buyers would thereby be denied the opportunity to pick the better diamonds from each category, they would nonetheless have the incentive to inspect each sight very carefully. Refusal to accept would signal that a sight was over-priced—but no more.

Suppose now that an in-or-out trading rule is added. The decision to refuse a sight now has much more serious ramifications. To be sure, a refusal could indicate that a particular sight is egregiously over-priced. More likely, however, it reflects a succession of bad experiences. It is a public declaration that de Beers is not to be trusted. In effect, a disaffected buyer announces that the expected net profits of dealing with de Beers under these constrained trading rules is negative.

Such an announcement has a chilling effect on the market. Buyers who were earlier prepared to make casual sight inspections are now advised that there are added trading hazards. Everyone is put on notice that a confidence has been violated and to inspect more carefully.

Put differently, the in-or-out trading rule is a way of encouraging buyers to regard the procurement of diamonds not as a series of independent trading events but as a long-term trading relation. If, overall, things can be expected to "average out," then it is not essential that an exact correspondence between payment made and value received be realized on each sight. In the face of systematic underrealizations of value, however, buyers will be induced to quit. If, as a consequence, the system is moved from a high to a low trust trading culture, then the costs of marketing diamonds increase. de Beers has strong incentives to avoid such an adverse outcome—whence, in a regime which combines all-or-none with in-or-out trading rules, will take care to present

16. *United States* v. *Loew's Inc.*, 371 U.S. 38 (1962).

sights such that legitimate expectations will be achieved. The combined rules thus infuse greater integrity of trade.

5.2. Economics of the Family

Transaction cost economics has been brought to bear on the economics of family organization in two respects: the one deals with family firms and productive relations; the other deals with "career marriages."

5.2.1. Family firms

Pollak's (1985) recent examination of families and households actually addresses a broader subject than family firms. I nevertheless focus these remarks on the family firm issue.

Pollak introduces his article with the following overview of the literature:

> The traditional economic theory of the household focuses exclusively on observable market behavior (i.e., demand for goods, supply of labor) treating the household as a "black box" identified only by its preference ordering. The "new home economics" takes a broader view, including not only market behavior but also such nonmarket phenomena as fertility, the education of children, and the allocation of time. The major analytic tool of the new home economics is Becker's household production model, which depicts the household as combining the time of household members with market goods to produce the outputs or "commodities" it ultimately desires.
>
> The new home economics ignores the internal organization and structure of families and households. Although this may surprise noneconomists who tend to believe that the internal organization and structure of an institution are likely to affect its behavior, economists find it natural. For the economist the most economical way to exploit the fundamental insight that production takes place within the household is to apply to households techniques developed for studying firms. Since neoclassical economics identifies firms with their technologies and assumes that firms operate efficiently and frictionlessly, it precludes any serious interest in the economizing properties of the internal structure and organization of firms. The new home economics, by carrying over this narrow neoclassical view from firms to households, thus fails to exploit fully the insight of the household production approach. . . . [By contrast,] the transaction cost approach which recognizes the significance of internal structure provides a broader and more useful view of the economic activity and behavior of the family. (1985, pp. 581–82)

Pollak then goes on to examine the strengths and limitations of the family in governance structure and technological respects and identifies the circumstances where family firms can be expected to enjoy a comparative advantage. The advantages of the firm are developed under four headings: incentives, monitoring, altruism, and loyalty. The main disadvantages of the family as a production unit are conflict spillover from nonproduction into production activities, a propensity to forgive inefficient or slack behavior, access to a restricted range of talents, and possible diseconomies of small scale. He con-

cludes that the strongest case for the family firm is "in low-trust environments (that is, in societies in which nonfamily members are not expected to perform honestly or reliably) and in sectors using relatively simple technologies" (1985, p. 593).

5.2.2. *Career marriages*

Career marriages of two kinds can be distinguished. One of these involves the marriage of a manager with a firm. The other involves cohabitation by two people, usually but not always of the opposite sex. The analysis here deals with the latter, but much of the argument carries over to marriages of manager and firm with minor changes.

I examine career marriages in the context of the contracting schema set out in Figure 3.1. Career being the entire focus, the parties are assumed to contract for marriage in a wholly calculative way.

Recall that node A corresponds to the condition where $k = 0$. Neither party in these circumstances makes career sacrifices in support of, or at the behest of, the other. This is strictly a marriage of convenience. Each party looks exclusively to his/her own career in deciding on whether to continue the marriage or split. If, for example, a promotion is offered in a distant city to one but not both, the marriage is severed and each goes his/her own way. Or if one job demands late hours or weekends and this interferes with the leisure time plans of the other, each seeks a more compatible mate. A wholly careerist orientation is thus determinative. Nothing being asked or given, there are no regrets upon termination.

The case where $k > 0$ is obviously the more interesting. Nodes B and C here describe the relevant outcomes.

A $k > 0$ condition is one in which one of the parties to the marriage is assumed to make career sacrifices in support of the other. Let X and Y be the parties, and assume that X subordinates his/her career for Y. Thus, X may help Y pay for his/her education by accepting a menial job that pays well but represents a distinctly inferior promotion track. Or X may agree to specialize in nonmarket transactions called "homemaking." Or X may agree to be available to Y as a companion. Not only are career sacrifices incurred, but X's homemaking and companionship skills may be imperfectly transferable if Y has idiosyncratic tastes.

Whatever the particulars, the salient fact is that X's future employment prospects are worsened by reason of career sacrifices made on behalf of Y.[17] The interesting question is: How will the life styles of such career marriages differ depending on whether Y offers a marriage safeguard to X or refuses one?

A node B outcome obtains if Y refuses (or is unable) to provide a safeguard to X. Under the assumption that contracts are struck in full awareness of the

17. This ignores the possibility that Y is a "celebrity" and that having been married to Y carries cachet. X then realizes an immediate status gain upon marriage. Career sacrifices by X can then be interpreted as "payment" for the status gain. But Y, under these circumstances, is the vulnerable party.

hazards, X will demand up-front pay for such circumstances. This is the condition to which Carol Channing had reference in the line "diamonds are a girl's best friend."

If, however, Y is willing and able to offer a safeguard, a node C outcome can be realized. Since X has better assurance under these circumstances that Y will not terminate the relation except for compelling reasons (because Y must pay a termination penalty), X's demands for current rewards (diamonds, dinner, travel, etc.) will be reduced.

This raises the question, however, of what form these safeguards can or do take. There are several possibilities, some of which are dependent on the prevailing legal rules.

Children provide a safeguard if the prevailing legal rules award custody to X and severely limit Y's visitation rights (place these rights under X's control). The award of other assets that Y is known to value also perform this function.

Dividing the property accumulated in the marriage and making alimony conditional on the magnitude of X's career sacrifice is another type of safeguard. In effect, such legal rules deny node B outcomes. If X is awarded wealth and income protection under the law, then Y will be deterred from terminating.

As with most deterrents, however, there are side-effects. Thus, Y can squander assets in contemplation of termination. And Y may refuse to work or flee if alimony payments are thought to be punitive.

A third possibility is to develop a reciprocal career dependency. This may not be easy, but it may be done (at some sacrifice, usually) in certain complementary career circumstances. A pair of dancers with a highly idiosyncratic style is one illustration. Lawyers with complementary specialties and idiosyncratic knowledge of a particular class of transactions (say, of a particular corporation) is another. An artist and his/her agent is a third possibility.

5.3. Corporate Finance

The Modigliani–Miller theorem that the cost of capital in a firm was independent of the proportion of debt and equity revolutionized modern corporate finance. It gave rise to an extensive literature in which a special rationale for debt in an otherwise equity-financed firm was sought. The first of these, unsurprisingly, was that debt had tax advantages over equity. But this was scarcely adequate. Further and more subtle reasons why debt would be used in preference to equity even in a tax-neutral world were also advanced. The leading rationales were: (1) debt could be used as a signal of differential business prospects (Ross, 1977); (2) debt could be used by entrepreneurs with limited resources who were faced with new investment opportunities and did not want to dilute their equity position, thereby to avoid sacrifice of incentive intensity (Jensen and Meckling, 1976); and (3) debt could serve as an incentive bonding device (Grossman and Hart, 1982).

The Modigliani–Miller theorem and each of the debt rationales referred to above treats capital as a composite and regards the firm as a production

Table 3.1.

Governance Feature	Financial Instrument	
	Debt	Equity
Contractual constraints	Numerous	Nil
Security	Pre-emptive	Residual claimant
Intrusion	Nil	Extensive

function. By contrast, transaction cost economics maintains that the asset characteristics of investment projects matter and furthermore distinguishes between debt and equity in terms of their governance structure attributes. The basic argument is this: the investment attributes of projects and the governance structure features of debt and equity need to be aligned in a discriminating way. The key governance structure differences between debt and equity are shown in Table 3.1.

The transaction cost approach maintains that some projects are easy to finance by debt and *ought to be financed by debt.* These are projects for which physical asset specificity is low to moderate. As asset specificity becomes great, however, the pre-emptive claims of the bondholders against the investment afford limited protection—because the assets in question have limited redeployability. Not only does the cost of debt financing therefore increase, but the benefits of closer oversight also grow. The upshot is that equity finance, which affords more intrusive oversight and involvement through the board of directors (and, in publicly held firms, permits share ownership to be concentrated), is the preferred financial instrument for projects where asset specificity is great. The argument is developed in Chapter 7.

5.4. The Modern Corporation

Transaction cost economics appeals to the business history literature for the record and description of organizational innovations.[18] The work of Alfred Chandler, Jr. (1962, 1977) has been especially instructive. Among the more notable developments have been the invention of the line and staff structure by the railroads in the mid-nineteenth century, the *selective* appearance of vertical integration (especially forward integration out of manufacturing into distribution) at the turn of the century, and the appearance in the 1920s and subsequent diffusion of the multidivisional structure.

Transaction cost economics maintains that these innovations are central to an understanding of the modern corporation. The study of such organizational innovations requires, however, that the details of internal organization be

18. Arrow observes that "truly among man's innovations, the use of organization to accomplish his ends is among both his greatest and earliest" (1971, p. 224). And Cole asserts that "if changes in business procedures and practices were patentable, the contributions of business change to the economic growth of the nation would be as widely recognized as the influence of mechanical innovations or the inflow of capital from abroad" (1968, pp. 61–62).

examined. That technological and monopoly conceptions of the corporation ruled in an earlier era is precisely because the details of internal organization were passed off as economically irrelevant.

From a transaction cost point of view, the main purpose of studying internal organization is to better understand the comparative efficacy of internal governance processes. What are the ramifications—for economizing on bounded rationality; for attenuating opportunism; for implementing a program of adaptive, sequential decisionmaking—of organizing the firm this way rather than that? The shift from the functionally organized (U-form) structure by large corporations that began in the 1920s is especially noteworthy.

The M-form innovation began as an effort to cope. Chandler's statement of the defects of the large U-form enterprise is pertinent:

> The inherent weakness in the centralized, functionally departmentalized operating company . . . became critical only when the administrative load on the senior executives increased to such an extent that they were unable to handle their entrepreneurial responsibilities efficiently. This situation arose when the operations of the enterprise became too complex and the problems of coordination, appraisal, and policy formulation too intricate for a small number of top officers to handle both long-run, entrepreneurial, and short-run operational administrative activities. (Chandler, 1962, pp. 382–83)

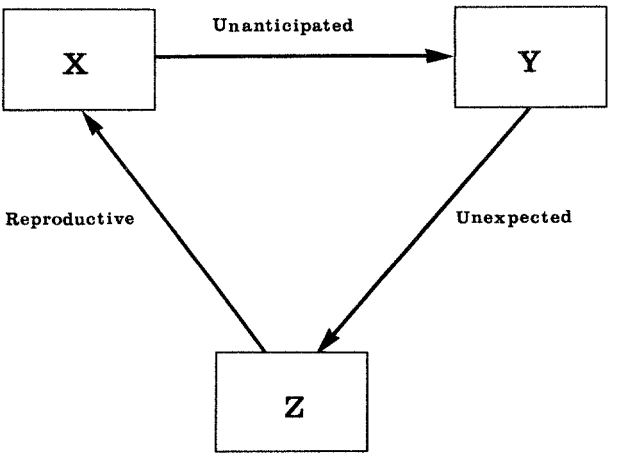
Bounds on rationality were evidently reached as the U-form structure labored under a communication overload. Moving to a decentralized structure relieved some of these strains.

But there was more to it than this. The M-form structure served not only to economize on bounded rationality, but it further served (in comparison with the U-form structure which it supplanted) to attenuate subgoal pursuit (reduce opportunism). This is because, as Chandler puts it, the M-form structure "clearly removed the executives responsible for the destiny of the entire enterprise from the more routine operational activities, and so gave them the time, information, and even psychological commitment for long-term planning and appraisal" (1966, p. 382).

The upshot is that the M-form innovation (X), which had mainly bounded rationality origins, also had unanticipated effects on corporate purpose (Y) by attenuating subgoal pursuit. Benefits of two kinds were thereby realized in the process.

There were still further unexpected consequences in store, moreover. Once the M-form organization had been perfected and extended from specialized lines of commerce (automobiles; chemicals) to manage diversified activities, it became clear that this structure could be used to support takeover of firms in which managerial discretion excesses were occurring (Z). A transfer of resources to higher valued purposes arguably obtains (Williamson, 1985b, pp. 319–22).

The spread of multidivisionalization through takeover thus yields the *reproductive link* that Elster notes is normally missing in most functional arguments in social science (1983, p. 58). The requisites of full functionalism are evidently satisfied.



X : M-form innovation
Y : Attenuated subgoal pursuit
Z : Takeover

Figure 3.6. Full functionalism.

Indeed, there is an additional process of spreading the M-form that ought also to be mentioned: mitosis. The large and diversified M-form structure may discover that the benefits associated with new activities or acquisitions do not continue indefinitely. Acquired components or diversified parts may therefore be divested. To the extent that these are spun-off or otherwise divested as discrete multidivisional units themselves, propagation through cell division may be said to exist. This quasi-biological process would also presumably qualify as a reproductive link and thereby contribute to successful functional explanation. Figure 3.6 summarizes the argument.

6. The Evidence

Transaction cost economics operates at a more microanalytic level of analysis than does orthodoxy. Whereas prices and quantities were thought to be the main if not the only relevant data in the orthodox scheme of things (Arrow, 1971, p. 180), transaction cost economics looks at the attributes of transactions and maintains that the details of organization matter. Additional data thus come under review.

Although the costs of such data collection can be great, resolution gains are frequently realized. Recent microanalytic studies in which transaction costs are featured are surveyed in Joskow, 1988 and Klein and Shelanski, 1995. To be sure, many empirical studies and tests of transaction cost economics are crude, yet the main implications are borne out and/or fare well in comparison with the leading alternatives. The crudeness to which I refer has two sources.

First, transaction cost theory and models are still very primitive. Only gross predictions are usually available. Secondly, severe measurement problems are posed. Both limitations will be mitigated as better models and better data become available.

Albeit real, current data limitations ought not to be exaggerated. Empirical researchers in transaction cost economics have had to collect their own data. They have resolved the trade-off of breadth (census reports; financial statistics) for depth (the microanalytics of contract and investment) mainly in favor of the latter. In the degree to which a subject becomes a science when it begins to develop its own data, this data switch is a commendable response.

7. Public Policy Ramifications

Transaction cost economics can be brought to bear on a wide variety of public policy issues. Although most of the applications have dealt with matters of microeconomic policy, the transaction cost economics perspective can also help to inform public policy toward stagflation.

7.1. Microeconomics

Microeconomic applications include regulation and antitrust. Consumer protection is another possibility.

7.1.1. Regulation/deregulation

Monopoly supply is efficient where economies of scale are large in relation to the size of the market. But, as Friedman laments, "There is unfortunately no good solution for technical monopoly. There is only a choice among three evils: private unregulated monopoly, private monopoly regulated by the state, and government operation" (1962, p. 128).

Friedman characterized private unregulated monopoly as an evil because he assumed that private monopoly ownership implied pricing on monopoly terms. As subsequently argued by Demsetz (1968b), Stigler (1968), and Posner (1972), however, a monopoly price outcome can be avoided by using ex ante bidding to award the monopoly franchise to the firm that offers to supply product on the best terms. Demsetz advances the franchise bidding for natural monopoly argument by stripping away "irrelevant complications"—such as equipment durability and uncertainty (1968b, p. 57). Stigler contends that "customers can auction off the right to sell electricity, using the state as an instrument to conduct the auction.... The auction ... consists of [franchise bids] to sell cheaply" (1968, p. 19). Posner agrees and furthermore holds that franchise bidding is an efficacious way by which to award and operate cable TV franchises.

Transaction cost economics recognizes merit in the argument but insists that both ex ante and ex post contracting features be examined. Only if

competition is efficacious at *both* stages does the franchise bidding argument go through. The attributes of the good or service to be franchised are crucial to the assessment. Specifically, if the good or service is to be supplied under conditions of uncertainty and if nontrivial investments in specific assets are involved, the efficacy of franchise bidding is highly problematic. Indeed, the implementation of a franchise bidding scheme under those circumstances essentially requires the progressive elaboration of an administrative apparatus that differs mainly in name rather than in kind from the sort associated with rate of return regulation.

This is not, however, to suggest that franchise bidding for goods or services supplied under decreasing cost conditions is never feasible or to imply that extant regulation or public ownership can never be supplanted by franchise bidding with net gains. Examples where gains are in prospect include local service airlines and, possibly, postal delivery. The winning bidder for each can be displaced without posing serious asset valuation problems, since the base plant (terminals, post office, warehouses, and so on) can be owned by the government, and other assets (planes, trucks, and the like) will have an active second-hand market. It is not, therefore, that franchise bidding is totally lacking in merit. On the contrary, it is a very imaginative proposal. Transaction cost economics maintains, however, that all contracting schemes—of which franchise bidding for natural monopoly is one—need to be examined micro-analytically and assessed in a comparative institutional manner. The recent examination of alternative modes for organizing electricity generation by Joskow and Schmalensee (1983) is illustrative.

7.1.2. Antitrust

The inhospitality tradition maintains the rebuttable presumption that nonstandard forms of contracting have monopoly purpose and effect. The firm-as-production function theory of economic organization likewise regards vertical integration skeptically. Integration that lacks technological purpose purportedly has monopoly origins [Bain, 1968, p. 381). The argument that "vertical integration loses its innocence if there is an appreciable degree of market power at even one stage of the production process" (Stigler, 1955, p. 183)—a 20 percent market share being the threshold above which market power is to be inferred (Stigler, 1955, p. 183)—is in this same spirit.

Transaction cost economics views integration differently. It maintains the rebuttable presumption that nonstandard forms of contracting, of which vertical integration is an extreme form, have the purpose and effect of economizing on transaction costs. It thus focuses on whether the transactions in question are supported by investments in specific assets. It furthermore examines monopoly purpose in the context of strategic behavior.[19]

19. Strategic behavior has reference to efforts by established firms to take up advance positions in relation to actual or potential rivals, to introduce contrived cost disparities, and/or respond punitively to new rivalry. Suffice it to observe here that strategic behavior is interesting only in an intertemporal context in which uncertainty and specific assets are featured.

Consider, in this connection, two stages of supply—which will be referred to generically as stages I and II (but for concreteness can be thought of as production and distribution). If the leading firms in a highly concentrated stage I were to integrate into an otherwise competitive stage II activity, the nonintegrated sector of the market may be so reduced that only a few firms of efficient size can service the stage II market. Then, entry would be deterred by the potential entrant's having to engage in small-numbers bargaining with those few nonintegrated stage II firms. Furthermore, the alternative of integrated entry will be less attractive because prospective stage I entrants that lack experience in stage II activity would incur higher capital and start-up costs were they to enter both stages themselves. If, instead, stages I and II were of low or moderate concentration, a firm entering either stage can expect to strike competitive bargains with either integrated or nonintegrated firms in the other stage, because no single integrated firm can enjoy a strategic advantage in such transactions, and because it is difficult for the integrated firms to collude. Except, therefore, where strategic considerations intrude—namely, in highly concentrated industries where entry is impeded—vertical integration will rarely pose an antitrust issue.

Whereas the original 1968 Guidelines reflected pre-transaction cost thinking and imposed severe limits on vertical integration (the vertical acquisition of a 6 percent firm by a 10 percent firm was above threshold), the revised Guidelines are much more permissive. The 1982 Guidelines are congruent with the policy implications of transaction cost economics in three respects. First, the 1982 Guidelines express concern over the competitive consequences of a vertical merger only if the acquired firm is operating in an industry in which the HHI exceeds 1800. The presumption is that nonintegrated stage I firms can satisfy their stage II requirements by negotiating competitive terms with stage II firms where the HHI is below 1800. The Guidelines thus focus exclusively on the monopolistic subset, which is congruent with transaction cost reasoning. Second, the anticompetitive concerns in the Guidelines regarding costs of capital, (contrived) scale diseconomies, and the use of vertical integration to evade rate regulation are all consonant with transaction cost reasoning. Finally, the Guidelines make express reference to the importance of asset specificity, although the analysis is less fully developed than it might be. Also, whereas the 1982 Guidelines make no provision for an economies defense, the 1984 Guidelines take this further step—which provision is especially important where asset specificity is demonstrably great.

7.2. Macroeconomics: Stagflation

Martin Weitzman's notable treatment of stagflation in his influential book *The Share Economy* mainly works out of a monopolistic competition framework. Weitzman augments the standard monopolistic competition apparatus, however, by distinguishing between redeployable and nonredeployable assets. Thus, he regards labor as redeployable while intermediate product is not: a "coalminer and a fruitpicker are infinitely closer substitutes than the products

they handle. Rolled sheet and I-beams . . . are virtually inconvertible in use" (Weitzman, 1984, p. 28). Unfortunately, this is a technological rather than a transactional distinction.

Such a technological view leads to a much different assessment of the contracting process than does a contractual view. Thus, whereas Weitzman regards labor market contracting as unique and flawed by rigidities, transaction cost economics maintains that labor markets and intermediate product markets are very similar and puts a different construction on rigidities. In particular, an examination of the governance needs of contract discloses that the full flexibility of wages and prices advocated by Weitzman would pose a serious threat to the integrity of contracts that are supported by durable investments in firm-specific assets. The lesson is that macroeconomics needs to come to terms with the study of contracting of a more microanalytic kind (Wachter and Williamson, 1978).

8. Conclusions

Friction, the economic counterpart for which is transaction costs, is pervasive in both physical and economic systems. Our understanding of complex economic organization awaits more concerted study of the sources and mitigation of friction. What is referred to herein as transaction cost economics merely records the beginnings of a response.

Refinements of several kinds are in prospect. One is that many of the insights of the transaction cost approach will be absorbed within the corpus of "extended" neoclassical analysis. The capacity of neoclassical economics to expand its boundaries is quite remarkable in this respect. Second, transaction cost arguments will be qualified to make allowance for process values such as fairness that now appear in a rather ad hoc way. (As Michelman [1967] has demonstrated, however, fairness and efficiency considerations converge when an extended view of contracting in its entirety is adopted. This insight is important and needs further development.) Third, numerous phenomena have yet to be brought under the lens of transaction cost reasoning. Recent experience suggests that new insights and new models are both in prospect. Fourth, a more carefully and fully developed theory of bureaucracy is greatly needed. Among other things, the powers and limits of alternative forms of internal organization with respect to reputation effects, internal due process, complex contingent rewards, auditing, and life cycle features need to be assessed. Finally, empirical research on transaction cost issues has been growing exponentially.