NBER WORKING PAPER SERIES

TRANSACTIONS COSTS AND INTERNAL LABOR MARKETS

Sherwin Rosen

Working Paper No. 2407

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 1987

Presented at the Conference Celebrating the 50th Anniversary of "The Nature of the Firm," Yale School of Organization and Management, May 14-16, 1987, forthcoming in The Journal of Law, Economics and Organization. The research reported here is part of the NBER's research program in Labor Studies. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

NBER Working Paper #2407
October 1987

Transaction Costs and Internal Labor Markets

ABSTRACT

The concept of transactions costs used by Coase in "The Nature of the Firm" is applied to the internal labor market of an organization. Under joint production it is shown that the number of transaction-specific prices necessary to decentralize labor allocations rises geometrically with the size of the work force. Complexity of calculation and costs of implementation constrains the possibilities for internal decentralization through a price mechanism and substitutes a more authoritarian system of allocation instead. These same issues of complexity and implementation costs limit the usefulness of agency theory as a conceptual framework for this problem. The analysis suggests that an internal labor market must be viewed in a more comprehensive framework of a personnel management system.

Sherwin Rosen
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637

Revised
October 10, 1987

TRANSACTIONS COSTS AND LABOR MARKETS

Sherwin Rosen*

University of Chicago

I. INTRODUCTION

Coase's first lecture reveals a surprising aversion toward mathematics. Curious, coming from one of the few economists who has a THEOREM named after him. In fact an easy case can be made that Ronald is responsible for two theorems, a lemma, and, according to some, an identity. THE Coase Theorem is at this point well beyond further discussion. The Second Theorem is the remarkable one on the time-consistent, subgame perfect equilibrium for a durable goods monopolist--the poor soul who is forced to either destroy some property or else act as a perfect competitor because it is impossible to commit now to actions that are not credible in the future except when monopoly power remains unexploited (Coase [1972]). The Lemma is not as well known, but should be. It is stated in some remarkable work with Fowler in 1935, the first known attempt to fit an intertemporal arbitrage condition, an Euler equation, to real data. Coase and Fowler [1935] were dubious about the rationality of Cobweb theory as an explanation for the pig cycle. Raising pigs happens to be a very specialized business. Breeders sell young pigs to feeders, who in turn sell them in the slaughter market after they have grown to the proper size. Coase and Fowler reasoned that easy money could be made in the first-stage transactions unless the

---

*I am indebted to Bengt Holmstrom, Edward Lazear, Oliver Williamson and the referee for comments and criticism of an initial draft and to the National Science Foundation for research support.

average market prices paid for young pigs reflected expectations about the price of pork some nine months later, the period in the 1930s (now it is six months) over which animals were held prior to slaughter, and they verified the hypothesis empirically in British data. The work is instantly recognized today as a version of the rational expectations hypothesis and acknowledged as such by Muth in his important paper on the subject.

Though many economists would sell their souls for a theorem, if not an inequality attached to their name, I suspect that Ronald would rather be identified with a LAW. This is not the place to speculate on why Laws in economics are so few, but they seem conspicuous by their scarcity. The law of demand is now a theorem about the Slutsky matrix, and most of the laws that are associated with specific name haven't fared well. Stigler's law was replaced, for a time, by Laffer's curve, the growth of Los Angeles annihilated Zipf's law, and the takeover and Japanese competition have not been kind to Gibrat's law. Walras and Engel have done better, but then only one of them concerns an empirical phenomenon. I hope Coase claims his law from his renewed interest in the subject at hand.

Many studies have calculated a four year average half-life of citations of articles in economics. Citations to the "The Nature of the Firm" show not only remarkable longevity but also an exceedingly rare decreasing hazard of mortality with age. No doubt this is due to the fundamental questions posed by the work and to the various meanings that can be attached to it. In reading it again I was struck by parallels with the literature of that time on the role of the price system in "spontaneously coordinating" economic activities, to use Hayek's felicitous expression, as compared to the heavy and inefficient hand of coordination through central planning. Coase takes the invisible hand as his point of departure and inquires into

the limits of market transactions as a coordinating mechanism. If markets are ideal coordinators, why should we ever observe any nonspontaneous, nonmarket coordination, as we appear to do within firms?

Coase argues that firms exist because some transactions internal to firms are less costly than similar transactions carried out in markets. The limits of the firm depend on cost comparisons at these margins.

Ultimately, these limits are determined by market competition among firms, including the market for corporate control. "Central planning" within firms is disciplined by competition them, so long as resources are free to move to their highest valued uses. As Alchian (1950) argued, firms making superior decisions gain control of more resources at the expense of the less efficient. It is the central role of competition and concern with more aggregate questions of supply and demand that probably accounts for why much of economic theory dispenses with the notion of the firm altogether; for example, general equilibrium theory uses only a very abstract notion of technology.

In what follows, I apply the theme of "The Nature of the Firm" to labor markets. Relation-specific exchange embodies the empirical content of transactions cost in modern industrial organization. Firm-specific human capital is a closely related concept. Section II reviews recent research showing that the costs of matching workers to firms and of assembling a team of workers are major components of these investments. Section III analyzes the nature of a decentralized market mechanism under these circumstances and shows that efficient allocations require a large number of transaction-specific prices. the costs and complexity of calculation and implementation make market decentralization impractical. The theme of complexity and a plethora of prices is pursued in section IV, in the context of principal and

agent theory. The emphasis here is on incentive rather than selection and allocation problems. Nonetheless, the main results so far share the same conceptual difficulty of excessive complexity and implementation costs. A broader approach which combines incentive, selection and allocation problems is stated in section V, within the context of the firm's personnel management policy of its internal labor market. Here changing selection and assignments of workers to positions over the work-life cycle interacts with performance incentives and worker capabilities. Conclusions appear in section VI.

II. TRANSACTIONS COSTS AND SPECIFIC CAPITAL

Coase did not define the empirical content of transactions costs in "The Nature of the Firm," nor tell us how to recognize them when we see them. Much progress has been made since then, especially by Becker [1964] and Williamson [1975], in identifying transactions costs with firm specific human and nonhuman capital. Shared investment costs requires sharing later returns and can lead to ex-post contract enforcement problems due to inefficient, opportunistic behavior. Several empirical observations in the labor market are consistent with the idea of specific human capital, especially the long-term attachments between workers and firms. The longest job of a typical white male worker persists for twenty-five years (Hall [1982]). Top level executives in major U.S. corporations are mostly "home grown," having spent thirty years or more with their firms in lesser positions before breaking into the top echelons (Murphy [1984]). Is is also otherwise difficult to understand the patterns of layoffs and employment variability among workers. Workers who have higher wages and greater job and firm-specific skills are less likely to be laid off.

So far, however, the magnitude of firm-specific human capital has eluded precise econometric measurement. The latest investigations of this problem by Altonji and Shakatko [1984]; Abraham and Farber [1987], Marshall and Zarkin [1985], and Topel [1987] suggest that most of the observed effects of firm-specific experience on earnings are due to selection. Highly paid employees have greater tenure with their firms but were also highly paid when first hired. These workers remain with their firms longer and exhibit greater firm-specific experience because their earnings were larger there to begin with. They were better matched to their jobs in the first instance. Workers who were not matched so well earned less and left their firm in search of greener pastures, thus exhibiting less firm-specific experience. When these selection effects are controlled statistically, it is found that the "true" firm specific experience effect on earnings is about the same as the general labor market experience effect, that is, the same as the general tendency for earnings to rise with age. Now match-specific effects certainly are a type of firm-specific capital, but of a slightly different nature than in the literature that derives inspiration from "The Nature of the Firm."

The measurement of physical capital-asset-specificity is perhaps easier, especially as it pertains to vertical integration. Joskow's [1987] recent study of the contractual relationships between electric utilities and coal suppliers is a good case in point. Nonetheless, there are ambiguities in defining the limits of the firm when asset specificities and transport cost-based rents are regulated by long-term contracts. Are these to be classified as market transactions, transactions internal to the extended family of the firm or what? Klein, Crawford and Alchian [1981] analyze many

examples where asset specificities are internalized by ownership. My favorite was the Hawaiian resort that purchased the adjacent golf course to avoid ex post bargaining costs and opportunism. However, that kind of reasoning won't go very far in accounting for why Yale University is vertically integrated with a splendid golf course (perhaps it is meant as an extra barrier for making tenure among golfers).

As pointed out by Shavell [1979], asset ownership dominates rental when the user's actions can substantially affect its resale or transfer value. Ownership internalizes conflicts of interest over maintenance and reckless use of equipment, and surely is the most important reason why most capital goods are owned outright by the firms that employ them. Some of the remaining cases of capital leasing can be understood on tax grounds (Scholes and Wolfson [1986]), but several defy analysis. Rentals of capital services are common in commercial real estate transactions, and greater tax advantages to wealthy individuals compared to businesses probably account for some of this, at least historically. Yet no such consideration applies to the separate ownership of sites and structures. Consider that the World Trade Center sits on rented land. Since those buildings are securely anchored to the bed rock below that part of Manhattan, it is difficult to conceive of more asset specificity than this. There is even more asset specificity in this than in Coase's observations on the contractual relations between GM and A.O. Smith in his third lecture. True, the lease on the land is very long term, running to 99 years. Still, the potential difficulties of renegotiation several years before the lease expires are well illustrated by what has happened in Hong Kong in recent times. These are more than rent-splitting and pure distribution problems, because the building owner can take actions, such as lack of maintenance, that directly

affects the value of the site. Common ownership of both site and structure would eliminate this problem. Why isn't it always observed?

III. LIMITS OF LABOR MARKET DECENTRALIZATION

Contractual difficulties arising from shared ownership of assets is an important case of a more general problem of devising decentralized pricing mechanisms under joint production. If there were no scale economies, transport costs nor economies of joint production, it is difficult to imagine why complete decentralization of labor markets would fail to achieve efficient allocations. Most workers would be, in some sense, self-employed. Coase provides a good example by reference to Stigler's [1951] discussion of British gun manufacture in the 18th century. Specialization and division of labor in allied trades was virtually complete when guns were manufactured on a small scale in a skilled craft system. Craftsmen were specialized by function: barrels, trigger mechanisms, stocks, sights, and so on. Others specialized in assembly, purchasing inputs from these specialists, producing the finished product and distributing it to customers. Most of these specialized transactions were carried out by market contracts, all within shouting distance of each other in a small area of Birmingham. Alfred Marshall [1930] analyzed this kind of system in his theory of external economies and locational concentration of specialized industries. Whitney's attempt to manufacture standardized guns on a large scale was unsuccessful, but his effort to achieve standardization and interchangeable parts altered gun manufacture forever. Gun-making was thereafter vertically integrated and many of the transactions that had previously been organized through the marketplace were coordinated by more authoritarian methods within firms.

Imagine how markets would have to be organized under these circumstances. A worker would own (or rent) a place in the assembly line, having purchased the rights from its previous owner. Its economic value would reside in the residual rights of contract, the profit gotten from purchasing intermediate products from adjacent upstream sellers and reselling the value-added units to adjacent downstream buyers. A decentralized contracting system confined to single quarters would be very difficult to manage because of the team aspects of the situation and the complicated interconnections of property rights they imply. Downstream workers, obliged to buy from an adjacent seller due to proximity and smaller transport costs, become very interested in the identity of that seller, because the volume and quality of work at each point affects the value of property rights of all others to whom it connects.

An exceedingly complicated contractual system, usually requiring side-payments among participants in the organization, is necessary to achieve efficiency in these circumstances. The number of prices necessary to manage it can be very large indeed. However, a simpler mechanism may be available: one person retains all residual rights, assembles the appropriate team of workers on a contractual basis, assigns them to their most productive positions in the firm, and monitors their work. The terms of these contracts must specify standards for the quality and quantity of work, as well as employment conditions regarding working hours and regularity of employment, these nonprice dimensions of contracts being necessary to internalize technological dependencies among workers. Financial terms of contracts are constrained by competition for workers in the labor market. Concentrating control in this way and establishing a wage system may be a less complicated way of achieving efficiency than designing and monitoring

an elaborate accounting system and calculating the individualized prices required by a decentralized internal transfer-pricing mechanism.

To illustrate the nature of the calculations needed, consider an organization where joint production entails complementarities of time spent with co-workers. Let x_i represent the output of worker i and let t_{ij} denote the time that i spends with j (the time that i spends alone is t_{ii}). There are n workers and the output of worker i is

$$x_i = F^i(t_{i1}, t_{i2}, \ldots, t_{in}) \quad \text{for } i = 1,2,\ldots,n. \tag{1}$$

The problem is to find an allocation of time $\{t_{ij}\}$ that maximizes total output in the organization, the sum of the x_i's, subject to two kinds of constraints. First, the time allocation of each worker must exhaust total time worked. Ignoring choice of total hours worked and normalizing it to 1.0 for each worker, there are n constraints of the form

$$1 = t_{i1} + t_{i2} + \ldots + t_{in}, \quad \text{for } i = 1,2,\ldots,n \tag{2}$$

In addition, the time that worker i desires to spend with worker j must equal the time that worker j desires to spend with worker i: there are $(n^2-n)/2$ constraints of the form

$$t_{ij} = t_{ji}, \quad \text{for } i \neq j. \tag{3}$$

First-order conditions for the efficient time allocation take the following form:

For t_{ii} we require

$$(4) \qquad F_i^i(t_{i1},t_{i2},\ldots,t_{in}) \leq \lambda_i, \quad \text{for } i = 1,2,\ldots,n$$

where λ_i is the multiplier on constraint (2) for worker $\underline{i}$. The equality is binding when $t_{\underline{ii}} > 0$, so $\lambda_{\underline{i}}$ has the interpretation of the shadow price of $\underline{i}$'s time. For $t_{\underline{ij}}$ and $t_{\underline{ji}}$ we need:

$$(5) \qquad F_j^i(t_{i1},t_{i2},\ldots,t_{in}) \leq \lambda_i + \beta_{ij} \quad \text{for } \underline{j} = 1,2,\ldots,n$$

$$F_i^j(t_{j1},t_{j2},\ldots,t_{jn}) \leq \lambda_j + \beta_{ji} \quad \text{for } \underline{i} = 1,2,\ldots,n$$

with strict equality whenever $t_{\underline{ij}} > 0$. Here $\beta_{\underline{ij}}$ is the multiplier associated with constraint (3) and $\beta_{\underline{ij}} = -\beta_{\underline{ji}}$. Since $\lambda_{\underline{i}}$ is the marginal product of own time, (4) and (5) together imply that if it is efficient for $\underline{i}$ and $\underline{j}$ to work together ($t_{\underline{ij}} = t_{\underline{ji}} > 0$) then

$$(6) \qquad \partial F^i/\partial t_{ij} + \partial F^j/\partial t_{ji} = \lambda_i + \lambda_j = \partial F^i/\partial t_{ii} + \partial F^j/\partial t_{jj}.$$

Equation (6) resembles the condition for efficient joint production of a "public good." The right hand side is the marginal cost of joint production for the pair of workers, the output foregone if both had spent their time alone rather than together. The left hand side is the marginal value of joint production, the sum of the incremental products of working together.

Conditions (4) - (6) have an important implication, that the decentralized price system that implements the efficient program is very complicated. The fact that (5) and (6) refer to pairs of workers means that the marginal product of a given worker's time is not equated across all workers to whom he is assigned. The time-price worker $\underline{i}$ spends with another worker

$\underline{k}$ is $\lambda_i + \beta_{\underline{ik}}$ and $\beta_{\underline{ij}} \neq \beta_{\underline{ik}}$ unless workers $\underline{j}$ and $\underline{k}$ are in some sense identical to each other. Hence a pricing system would have to use different prices for time charged in each possible pairing. This point is related to the problem of establishing prices in network systems such as landing rights at airports and other assignment problems (Koopmans and Beckmann [1957]).

Let $w_{\underline{ij}}$ be the unit price that worker $\underline{i}$ charges worker $\underline{j}$ per unit time, with $w_{\underline{ij}} = -w_{\underline{ji}}$, so $w_{\underline{ji}}$ is the unit price $\underline{j}$ must pay to $\underline{i}$ if $w_{\underline{ij}}$ is positive (or the price that $\underline{j}$ charges $\underline{i}$ per unit time if $w_{\underline{ji}}$ is negative). Taking output as numeraire, the decentralized solution is achieved by letting each worker act as a residual income recipient, selling own output to the owner of the firm at price 1.0 and charging each co-worker $w_{\underline{ij}}$ per unit of time spent with each. Worker $\underline{i}$ chooses $\{t_{\underline{ij}}\}$ to maximize

$$F^i(t_{1i},t_{2i},\ldots,t_{ni}) + \sum_{i\neq j} w_{ij}t_{ij} \tag{7}$$

subject to (2). The first order condition is

$$-\partial F^i/\partial t_{ii} + \partial F^i/\partial t_{ij} + w_{ij} \leq 0, \quad \text{for } i \neq j. \tag{8}$$

This results in the efficient solution so long as $\beta_{\underline{ij}} = w_{\underline{ij}}$, that is, so long as the proper price of time (possibly so large that $\underline{i}$ and $\underline{j}$ do not work together) is found for each pair. Since there are $(n^2 - n)/2$ independent shadow values of time, the number of prices necessary to achieve efficiency increases with the square of the number of workers. It would increase even more if triples and higher orders of joint production were considered. Moreover, to calculate and implement this solution requires full knowledge

of the underlying technology and productivity of team members in the first place.

If that knowledge is possessed by a specialist, an authoritarian system whereby the specialist-manager allocates workers to each other and monitors their activities may be less expensive to implement than an internal price system. Errors in prices can be more costly than errors in quantities (Weitzman [1974]). For example, complementarities may be so large that the optimal t_{ij}'s are easy to calculate, whereas small errors in setting the prices w_{ij}'s could lead to serious misallocations of time among individual workers. Furthermore, agreeing on a price can be time-consuming and divert time and energy away from production even when it is clear that trade should take place. For if exact valuations are private information then traders have incentives to argue over the distribution of gains from trade. Of course elements of these very same problems arise in interfirm as well as intrafirm transactions. Nevertheless, direct team interactions are far less important in interfirm transactions and they are more easily regulated by contractural arrangements, the monitoring of output quality, and by market competition among alternative sources of supply. The close-quarter interactions of workers and the transport cost savings they imply limit the degree of substitution and competition from outside alternatives. External labor market competition disciplines a firm's internal labor market with respect to overall wages and working conditions, but leaves some slack at the micro-transactions level of precise worker interactions.

Such a system is observed in our own backyard, in the organization of the education industry. In modern educational systems, the price mechanism is used largely to allocate students and teachers among schools, and even then it is incompletely used for this purpose: nonprice considerations

play an important role in these allocations. It is used hardly at all to allocate students to courses and to teachers within schools. Gaining admission and paying tuition entitles a student to fish among a broad range of courses. Committees and other collective bodies determine requirements, course sequencing, class size and other matters of internal allocations. Transfer prices are seldom used.

It was not always so. The original universities were collections of individual teacher-entrepreneurs, and fees were determined by bargaining and haggling with individual students (Rashdall [1895]). As universities emerged out of these crude beginnings, two-part pricing schemes were adopted. Students paid lump-sum tuition charges to enter and a marginal payment to specific teachers in courses of their choice. This is the system that Adam Smith advocated, on incentive and agency grounds. But as far as one can tell, all teachers in the same university charged the same unit prices, whereas efficient allocations of students to teachers almost certainly require different prices for different teachers, as well as price differentials among students within each course. Two-part pricing was entirely abandoned in the twentieth century and replaced with one-part salary and tuition pricing, probably because the bundling and information aspects of modern formal education made it cost effective to ration by queues, prerequisites and requirements rather than by individually tailored prices (see Rosen [1987] for further elaboration).

IV. DECENTRALIZATION AND AGENCY

If the number of prices necessary to decentralize a complex interactive organization increases multiplicatively with size, then the amount and cost of monitoring required to achieve efficiency also must grow with size.

Information becomes garbled as it passes through longer chains, and information channels become congested as chains-of-command lengthen (Williamson [1967]). It is the balancing of joint-production and scale economies against increasing costs of control associated with nonmarket transactions costs that determines organization size in traditional theory.

The avoidance of monitoring cost preoccupies the modern literature on agency theory. The main question posed is: Can penalty and reward systems be found that result in self-enforcing contracts? If so then internal decentralization may be efficient and the size of firms could be very large indeed. A fundamental result proves that multipart pricing is necessary to induce an agent to behave in the interests of the principal. This is the bonding scheme analyzed by Becker and Stigler [1974]. The idea is straightforward and rests on the proposition that an agent behaves honestly if confronted by a scheme that makes such behavior consistent with self interest. Evidently the scheme must either reward good behavior or punish bad behavior (malfeasance).

Considerations of labor market equilibrium dictate the penalty mode rather than the reward mode. For if the agent is given extra monetary rewards for good behavior, the expected utility from holding the job exceeds that available from alternatives and the supply of job applicants exceeds the number of available positions. On the other hand, if a worker posts bond money "up front" and the bond can be seized by the firm if malfeasance occurs, honest behavior is elicited by paying a market wage premium equivalent to interest on the bond, with the bond itself returned at the end of the contract. This bond-interest-principle scheme equalizes workers' expected returns among jobs and achieves job-market clearing. An important modification of the argument allows workers to post bond by investing in the

firm; by working at a wage less than marginal product in the early years of a career, and receiving the return at older ages in the form of wage payments in excess of productivity (Lazear [1979]). Another modification with risk aversion (Mirrlees [1974]) also favors the penalty mode because potential monetary losses reduce utility by more than equal monetary gains increase utility.

Potential penalties must increase with the agent's perceived returns to malfeasance to elicit honest behavior in bonding schemes. The temptation toward malfeasance is decreasing in the extent of monitoring and detection activity by the firm as well as in the size of the bond to be lost if malfeasance is detected. It follows that monitoring and the size of the bond are inversely related in bonding schemes. But since monitoring uses real resources (monitors must be hired and taken out of some other productive use of labor), whereas bonds do not, monitoring resources can be driven to zero as the bond increases without limit. The scheme is completely self-enforcing in the limit. For example, penalizing double-parking offenses by execution would reduce the incidence of double-parking to miniscule proportions and very little police time would have to be spent in ticketing offenders. Even apart from the time-inconsistent (incredible) nature of this extreme example, these limiting results are mainly of academic interest. For as the bond grows in size the principal is more likely to find malfeasant behavior when it isn't there. This type-II error is itself a manifestation of malfeasance of another kind for large bonds increase the propensity for the firm to find the employee "guilty" and seize the bond. Hence it is not feasible to eliminate monitoring, and the optimum scheme must involve both penalties and monitoring.

Possibilities for malfeasance by multiple agents in joint production require mutual monitoring and "double" bonding by all participants. This problem has not been completely analyzed, though some interesting work has appeared on the role of reputations in serving as bonds; and agency considerations have been introduced into the analysis of trade unionism, where the union serves as a worker's agent in dealing with the firm. An earlier approach derives from Marshall's critique of sharecropping, where rewards are stipulated as shares of gross revenue rather than of net profit. Incentives by sharecroppers and landlords are misaligned because both receive only a fraction of their social marginal product in deciding how much labor and effort to supply to the venture. Marginal private return falls short of marginal social return and effort is too small (Johnson [1951], Cheung [1969]).

In a multiple sharing arrangement, the socially efficient production outcome occurs only if the marginal share is unity for each party: each receives full marginal product in equilibrium (Groves [1973]). Various mechanisms have been studied to implement the efficient solution; including "budget breaking" (Holmstrom [1982]), double-bonds (Kennan [1979]), and trigger-strategies in repeated games (Radner [1981]), though little empirical research has studied the frequency with which any of them are observed in practice. Since simple sharecropping systems have been historically important in the organization of agriculture and similar institutions are commonly observed in contingent fees for lawyers, the division of reward among doctors (Gaynor and Pauly [1987]) and lawyers (Gilson and Mnookin [1985]) in group practice, book royalty arrangements, rewards to actors, musicians and so forth, the survivor principle suggests that the efficiency losses from these schemes must have been kept at tolerable proportions. The

simplest hypothesis is that joint monitoring and the adverse effects of shirking on reputations and future business dealings play important roles in resolving these conflicts of interest.

Another approach to the principal-agent problem generalizes decentralized output-reward systems to include considerations of risk sharing (Holmstrom [1979]). The problem is set up to investigate the consequences of hidden actions of the agent. The principal cannot observe the agent's action, but can observe the output that is the result of these actions. There cannot be a one-to-one correspondence between output and action or else the principal could infer actions perfectly and the problem is trivial. So output is a mixture of random effects and unobserved actions. If output is large, the principal cannot tell if the agent worked very hard or was very lucky. Similarly, a small output could have been due to bad luck rather than shirking. The worker is risk averse and prefers certain income to risky income, but observability constraints make it impossible to separate insurance from incentives. Paying a strict linear piece rate gives the agent proper incentives to expend effort because the agent realizes the full social product of effort, but at the cost of exposure to excessive risk. Paying a guaranteed wage provides full insurance, but does not provide any incentive to work.

The solution is a compromise between these two opposing forces. The earliest treatments (Stiglitz [1975]) analyzed two-part tariff solutions, where the principal guarantees the agent a minimum compensation for insurance reasons and a percentage of revenues to provide incentives to work hard. The proportion of pay in each part depends in an obvious way on the extent of risk, the elasticity of output with respect to effort, and the degree of risk aversion. However, when the problem is generalized to allow

the form of the payment schedule to be endogeneously determined, the solution is extremely complicated: payment need not even be everywhere increasing in output (Grossman and Hart [1983]).

The complicated payment schedules predicted by theory is an embarrassment of riches and another manifestation of "too many prices:" the schemes we observe, such as salesperson's commissions and contingency fees in legal practice, have very few parameters. These problems would imply complete decentralization and simple linear transfer prices were it not for the presence of risk aversion, so there is a sense in which risk aversion and insurance elements lead the theory astray. One can be properly skeptical that risk aversion and the precise form of preferences are such an important part of the problem. After all, a great virtue of a price system is that it works when utility and production functions are completely private information. Could it be that such simple schemes are observed because they are robust to varieties of preferences? Holmstrom and Milgrom [1987] have recently introduced intertemporal arbitrage considerations to enforce linearity onto the optimal scheme. This is an interesting idea, but the results still depend on special assumptions about risk aversion. None of this theory extends in any obvious way to problems involving joint production among several agents. Furthermore, the analysis assumes that principals possess complete information about preferences of others, and is hardly decentralized in that sense.

V. INTERNAL LABOR MARKETS

I have followed Coase [1937] and Alchian and Demsetz [1972] in arguing that the expense of implementing quasi-market decentralization

within firms forces analysis on the role of performance monitoring in understanding organizational structure. The interactions of personnel within organizations are too complicated to be completely decentralized through a price mechanism. Indeed, if this were not the case then Coase's argument implies that the firm should not exist. This theme is consistent with Williamson's [1975,1985] criticism of the textbook association of firms as production functions and his idea of a governance structure. The firm's observed production and cost functions are the outcomes of the interaction between production technology, personnel policy, management, and institutional rules and design.

Considerations of the long term goals and survival of the organization lend additional credence to this view. Since there is substantial earnings growth over the life cycle and since most job turnover occurs early in the working life-cycle, a large fraction of a person's life-cycle earnings is generated over the course of a career with one firm. Organizational complexity arises from the intertemporal aspects of personnel management systems. Organization dynamics cannot be separated from internal job mobility among overlapping generations of workers and management. All organizations require specialization and division of labor among their members, but job assignments systematically change over a person's tenure with the firm. Institutional memory, specific knowledge, skills and responsibilities are constantly being transferred from old to young.

The flow and throughput of personnel through positions in the firm can be thought of as an "internal labor market." A very good example is provided by the officer corps in the military, where all participants begin at a the lowest rank and either move up to higher positions of authority and command or leave for alternative employments outside the military. Most

organizations are more complicated than this because lateral entry and exit accur at many points, not simply at one point. Still, most follow a hierarchical design in which ultimate control is concentrated at the top and diffused through the ranks by horizontal and vertical linkages to middle and lower level management and to production. In large organizations, it is especially important to assign the most capable and energetic people to top level positions because top-level decisions percolate through the organization and have much larger effects on organizational productivity than lower level decisions do.

Top level decisions have multiplicative effects on productivity in management technologies where authority is limited by a span of control and where monitoring resources are partially economized through lengthy chains-of-command. These multiplicative effects imply that more capable top-level decision-making can have enormous effects on the organization and imply that the socially efficient assignment of personnel to positions is hierarchical in ability. The most capable people should control the most resources and direct the largest organizations. Less capable and less energetic people assigned to lower level positions in large firms or higher level positions in smaller firms. The interaction of talent and scale can support extremely large salaries for top level managers of large firms on marginal productivity grounds alone (Rosen [1982]), consistent with empirical findings that top executive compensation is systematically increasing with firm size (Murphy [1984], Kostiuk [1985]) as well as with profitability.

Monitoring, testing, and performance evaluation take on special significance under these circumstances. Resources must be continually devoted to designing career tracks and to grading, sorting and assigning workers to their proper positions in the organization. Employees are not

passive by-standers in this process because their incomes and status depend on how they are graded. The economics of this combined design and incentive process has begun to be analyzed in the literature on tournaments (Lazear and Rosen [1981], in which the firm optimizes its testing, selection procedures and wage structure against the competitive efforts of workers to affect their scores, elevate their classifications and achieve higher ranking positions. The ordinal quality of this kind of competition follows from the inherent ordering properties of tests and peer comparison when direct output measures are difficult to devise. Ordered or relative performance evaluation also has certain optimality properties in the presence of risk aversion: it eliminates extraneous variance due to measurement error that is common to all participants (Holmstrom [1982], Green and Stokey [1983], Nalebuff and Stiglitz [1983]).

Sequential statistical decisions that rank and order contestants are inherent in the intergenerational dynamics of organizations and lead to a theory of promotions through the ranks as an important motivator of the organization's members. Performance incentives are provided by the wage differentials between hierarchical ranks. Top ranking prizes (wages) take on special significance in this kind of competition, for they must rise more than in proportion with rank to maintain performance incentives among those competing for the highest level positions (Rosen [1986]). At early stages of a career a person's performance incentives are propelled by a kind of "option" value, the possibility of achieving not only the next highest position, but all possible positions higher than that. As a successful contestant progresses through the hierarchy and climbs higher in rank and authority, there are fewer places left to attain. The option value falls

with rank because there is less distance to travel. Increasing the difference in wages among the top-most ranking positions maintains incentives by substituting for the option value that propelled performance incentives at lower ranks. In this sense wage structures among top executive positions reflect both the productivity of top level managers and the productivity induced by the attempts of lower ranking employees to climb higher.

A problem inherent in performance evaluation and ability testing has received increasing attention in the literature. Since grading, evaluation and promotion decisions are made by higher level committees and supervisors, contestants have incentives to increase their scores by exerting unproductive "influence" on the examiners (Milgrom [1987]). For example, in relative performance evaluations, there may be gains from unproductive activities that degrade the ranking of competitors and make a contestant look better than others (Lazear [1986]). These adverse "gaming" incentives by contestants apply to any evaluation system (Baker [1987], Breton and Wintrobe [1986]) and help to understand some of the bureaucratic procedures adopted by organizations to control them. These bureaucratic costs are properly considered as transactions costs of nonmarket allocations within firms and may ultimately help define the limits of the firm.

VI. CONCLUSION

I have argued that the competitive price mechanism necessary to decentralize a complex interacting organization with indivisibilities and joint production is very complicated. So much information and preknowledge is required that more authoritarian "planning" mechanisms are likely to economize on transactions costs within firms. With respect to labor re-

sources, these allocation and contracting problems certainly involve firm-specific human capital. However, much of this appears to arise in the context of assembling a coherent work force and productive team within the firm, collecting and processing information on team members' talents and assigning them to their proper niche in the organization, and transferring productive knowledge between older and younger members of the organization.

Incentives, testing, career assignments and rewards must be analyzed in the context of a dynamic personnel system. Incentives and reward structures cannot be disassociated from testing, personnel assignment and labor turnover questions in such a system. In combining all of these functions, personnel policies are likely to be inefficient at some margins separately, though they may achieve reasonably good compromises among all goals considered together. Looking at these systems or internal-labor-market aspects of personnel management helps to understand some of the bureaucratic tendencies in organizations as controls on members' attempts to unproductively manipulate the system to personal advantage. Obviously much work remains to be done in this area, but if successful it will improve our understanding of the limits of firms and the limits of markets.

REFERENCES

Abraham, Katharine G. and Farber, Henry S. "Match Quality, Seniority and Earnings," American Economic Review 77 (June 1987): 278-97.

Alchian, Armen A. "Uncertainty, Evolution and Economic Theory," J.P.E. 58 (June, 1950): 211-221.

_____ and Demsetz, Harold. "Production, Information Costs, and Economic Organization," Amer.Econ.Rev. 62 (Dec., 1972): 777-795.

Altonji, Joe and Shakotko, Robert. "Do Wages Rise with Job Seniority?" Columbia University, July, 1984.

Baker, George P. "Monitoring Costs and Compensation Structure," Harvard Business School, May, 1987.

Becker, Gary S. The Theory of Human Capital: A Theoretical and Empirical Analysis. New York: Columbia Univ. Press, 1964.

_____ and Stigler, George J. "Law Enforcement, Malfeasance, and Compensation of Enforcers," J.of Legal Stud. 3 (Jan., 1974): 1-18.

Breton, Albert and Wintrobe, Ronald "The Bureaucracy of Murder Revisited," J.P.E. 94 (Oct., 1986): 905-926.

Cheung, Steven N. The Theory of Share Tenancy. Chicago: University of Chicago Press, 1969.

Coase, Ronald. "The Nature of the Firm," Economica (1937, no. 4): 386-405.

_____. "Durability and Monopoly," J. Law and Econ. 15 (April 1972): 143-49.

_____ and Fowler, R.H. "Bacon Production and the Pig Cycle in Great Britain," Economica (1935): 142-67.

Gaynor, Martin and Pauly, Mark. "Alternative Compensation Arrangements and Productive Efficiency in Partnerships: Evidence from Medical Group Practice," NBER, Feb., 1987.

Gilson, Ronald J. and Mnookin, Robert H. "Sharing Among the Human Capitalists: An Economic Inquiry into the Corporate Law Firm and How Partners Split Profits," Stanford Law Review 37 (January 1985): 313-97.

Green, Jerry and Stokey, Nancy. "A Comparison of Tournaments and Contracts," J.P.E. 91 (June, 1983): 349-65.

Grossman, Sanford J. and Hart, Oliver. "An Analysis of the Principal-Agent Problem," Econometrica 51 (Jan., 1983): 7-45.

Groves, Theodore. "Incentives in Teams," Econometrica 41 (July, 1973): 617-32.

Hall, Robert E. "The Importance of Lifetime Jobs in the U.S. Economy," Amer. Econ. Rev. 72 (Sept., 1982): 716-27.

Holmstrom, Bengt. "Moral Hazard in Teams," Bell J. Econ. 13 (Autumn, 1982): 324-40.

_____. "Moral Hazard and Observability," Bell J. Econ. 10 (Spring, 1979): 74-91.

_____ and Milgrom, Paul. "Aggregation and Linearity in the Provision of Intertemporal Incentives," Econometrica 55 (March, 1987): 303-29.

Johnson, D. Gale. "Resource Allocation under Share Contract," J.P.E. 68 (April, 1950): 111-23.

Joskow, Paul L. "Contract Duration and Relation-Specific Investments: Empirical Evidence from Coal Markets," Am. Econ. Rev. 77 (March, 1987): 168-85.

Kennan, John. "Bonding and Enforcement of Labor Contracts," Econ. Letters 1979.

Klein, Benjamin, Crawford, Robert G. and Alchian, Armen A. "Vertical Integration, Appropriable Rents and the Competitive Contracting Process," J.P.E. 89 (Aug., 1981): 615-641.

Koopmans, Tjalling and Beckmann, Martin. "Assignment Problems and the Location of Economic Activities," Econometrica 25 (Jan., 1957): 53-76.

Kostiuk, Peter. "Firm Organization and Compensation of Corporate Executives," unpublished Ph.D. dissertation, Univ. of Chicago, 1985.

Lazear, Edward P. "Pay Equality and Industrial Politics," Hoover Institution, 1986.

_____. "Why Is There Mandatory Retirement?" J.P.E. 87 (Dec. 1979): 1261-84.

_____ and Rosen, Sherwin. "Rank Order Tournaments as Optimum Labor Contracts," J.P.E. 89 (Oct., 1981): 841-64.

Marshall, Alfred. Principles of Economics (8th ed.). London: Macmillan, 1930.

Marshall, Robert C. and Zarkin, Gary A., "The Effects of Job Tenure on Wage Offers," Duke University, 1985.

Milgrom, Paul. "Employment Contracts, Influence Activities and Efficient Organization Design," Univ. of California, Berkeley, 1987.

Mirrlees, James A. "The Optimum Structure of Incentives and Authority within an Organization," Bell J. Econ. 7 (Spring, 1976): 105-31.

Murphy, Kevin J. "Ability, Performance and Compensation: A Theoretical and Empirical Investigation of Managerial Compensation," unpublished Ph.D. dissertation, Univ. of Chicago, 1984.

Nalebuff, Barry J. and Stiglitz, Joseph E. "Prizes and Incentives: Toward a General Theory of Compensation and Competition," Bell J. Econ. 14 (Spring, 1983): 21-43.

O'Keeffe, Mary, Viscusi, W. Kip, and Zeckhauser, Richard J. "Economic Contests: Comparative Reward Schemes," J. Labor Econ. 2 (Jan., 1984): 27-56.

Radner, Roy. "Monitoring Cooperative Agreements in a Repeated Principal-Agent Relationship," Econometrica 49 (Sept., 1981): 1127-1148.

Rashdall, Hastings. The Universities in Europe in the Middle Ages. Oxford: Oxford Univ. Press, 1895.

Rosen, Sherwin. "Authority, Control and the Distribution of Earnings," Bell J. Econ. 13 (Oct., 1982): 311-23.

_____. "Prizes and Incentives in Elimination Tournaments," Amer. Econ. Rev. 76 (Sept., 1986): 701-15.

_____. "Some Economics of Teaching," J. of Labor Econ. 1987 (forthcoming).

Scholes, Myron S. and Wolfson, Mark A. "Taxes and Organization Theory," Stanford U., Sept., 1986.

Shavell, Steven. "Risk Sharing and Incentives in the Principal and Agent Relationship," Bell J. Econ. 10 (Spring, 1979): 55-73.

Stigler, George J. "The Division of Labor is Limited by the Extent of the Market," J.Polit.Econ. 59 (Jan., 1951): 185-93.

Stiglitz, Joseph E. "Incentives, Risk and Information: Notes Toward a Theory of Hierarchy," Bell J. Econ. 6 (Autumn, 1975): 552-79.

Topel, Robert, "Job Mobility and Earnings Growth: A Reinterpretation of Human Capital Earnings Functions," in R.E. Ehrenberg (ed.), Research in Labor Economics, Greenwich: JAI Press, 1987.

Weitzman, Martin. "Prices Versus Quantities," Review of Economic Studies, 41 (Oct. 1974): 477-91.

Williamson, Oliver. Markets and Hierarchies:Analysis and Antitrust Implications, New York: Free Press, 1975.

_____. The Economic Institutions of Capitalism. New York: Free Press, 1985.

_____. "Hierarchical Control and Optimum Firm Size," J.P.E. 75 (April, 1967).