The Economics of Internal Labor Markets

MICHAEL L. WACHTER and RANDALL D. WRIGHT*

Our essay focuses on the economics of long-term contractual relationships between a firm and its employees, referred to as the ınternal labor market. We review the economics lıterature on match-specific ınvestments, risk aversion, asymmetrıc information, and transactıon costs. We argue that an ıntegrated treatment of all four factors ıs needed ın order to apply implicit contract theory to ınternal labor markets. Integratıng the topics also hıghlights the tradeoffs created among these factors. Our dıscussion stresses contract enforcement mechanısms, including self-enforcing contracts and third-party enforcement.

THE INTERNAL LABOR MARKET (ILM) consists of a set of explicit or implicit, more or less long-term agreements between a firm and its workers.[1] These agreements include implicit and explicit rules governing wages, hours of work, promotion opportunities, and grievance procedures. The manner in which the agreements are to be enforced, including self-enforcement and third-party enforcement mechanisms, are delineated as well. The terms of the contractual relationship also may be contingent on exogenous future events, such as changes in the firm's product market conditions or changes in the macro economy.

* The authors' affilıatıons are, respectıvely, Department of Economıcs and Institute for Law and Economics, Universıty of Pennsylvania, and Department of Economics, Unıversıty of Pennsylvania and Hoover Instıtution, Stanford Universıty. Valuable research support was provided by the Institute for Law and Economics of the Unıversity of Pennsylvania. We thank Costas Azarıadis, Harold Cole, Ronald Ehrenberg, Morris Kleiner, Ed Nosal, John Pencavel, Sherwin Rosen, and Mahmood Zaıdi for their helpful comments

[1] Long-term employment relationships are empırıcally ımportant. Hall (1982), for example, finds that the median completed tenure for workers ın the U S ıs 7.7 years, and that 28 per cent of workers are currently ın ȷobs that wıll last at least 20 years. Much ȷob turnover occurs among younger workers before they establısh long-term attachment (see Mincer and Jovanovıc, 1981). Furthermore, Feldsteın (1976) and Lılien (1980) find that over 70 per cent of layoffs are temporary, ındıcating that short-term separatıons usually do not end the long-run employment arrangement.

INDUSTRIAL RELATIONS, Vol 29, No 2 (Sprıng 1990) © 1990 Regents of the Unıversıty of Calıfornıa
0019/8676/90/525/240/$10.00

The analysis of internal labor markets began during the fifties when Kerr, Dunlop, and others developed the idea that the textbook market of supply and demand analysis could only accurately describe the external market for new hires, or the internal market of a few industries, such as agricultural labor and construction.[2] These scholars also first described the institutional realities of internal markets, their implications for the overall economy, and the "ports-of-entry" through which the external market influenced the internal markets. The resulting models were interdisciplinary, incorporating organizational behavior and legal aspects of the markets.[3]

This pioneering literature did not explicitly integrate the ILM into the neoclassical economic model. Consequently, it was sometimes assumed that a primary effect of the ILM was to reduce the importance of economic forces such as optimizing behavior. The early pioneers did not, however, take an anti-efficiency approach. Instead, they primarily attacked the uncritical application of the textbook model of supply and demand to the ILM.

The efficiency aspects of the ILM were first explicitly stressed in the seventies. Doeringer and Piore (1971) made the initial steps in developing some areas of compatibility between the ILM and the neoclassical model. Williamson, Wachter, and Harris (1975) and Okun (1981), among others, applied the developing contract literature to the ILM and consciously stressed the efficiency aspects of the ILM. Freeman and Medoff (1984) placed the role and impact of labor unions in a neoclassical perspective.

More recent economics literature has emphasized the efficiency theme in a more formal theoretical framework designed to isolate the central behavioral features of the ILM. Each potential behavioral assumption is separately modeled, in an effort to test the extent to which any single factor can provide a simple theoretical explanation of the employment relationship (see recent reviews by Flanagan [1984a], Rosen [1985], and Parsons [1986c]). This labor contracts literature has made great strides in developing a theoretical structure for the ILM, but the results remain incomplete. While factors such as risk aversion and specific training can explain important

[2] See, for example, Dunlop (1958) and Kerr (1954) and their recent retrospective reviews, Dunlop (1988) and Kerr (1988) This literature is extensively reviewed in Doeringer and Piore (1971).

[3] In reviewing the early ILM literature, Dunlop (1988, p. 50) noted, "An understanding of labor markets and compensation requires a recognition that the work place is a social organization, at least informally, and that labor markets take on significant social characteristics that do not characterize commodity and financial markets and that are not readily encapsulated in ordinary demand and supply analysis." The interdisciplinary nature of the ILM literature is evident in leading casebooks and monographs. In labor law, see for example, Cox, Box, and Gorman (1986); and in labor relations, Kochan, Katz, and McKersie (1986).

attributes of the ILM, no single element can do all of the work itself. As a consequence, this theoretical literature has not provided an integrated model of the ILM that might be useful to labor relations specialists and practitioners.

In this paper, we survey the diverse aspects of the current labor contracts literature in an attempt to bridge the separate strands of the theoretical literature and the more traditional interdisciplinary approach. We identify four factors—match-specific investments, risk aversion, asymmetric information, and transaction costs—which must be brought together to explain the full range of institutional rules in the ILM. We argue that the main contribution economic analysis can make to an interdisciplinary vision of the ILM is to identify how contract rules serve the optimizing goals of the firm and its workers. In moving toward a more integrated view of the contract that underlies the employment relationship, we focus on the enforcement aspects of the relationship. Enforcement problems are complex because the contract is designed to be long term but is also often incomplete. Nevertheless, these problems suggest ways in which the theoretical model can be applied to current issues.

Our main concern is with efficiency aspects of the ILM, and we spend less time on distributional issues. One of the strengths of the economic approach is that it allows the two to be separated. Thus, within the ILM there will generally exist a set of contractual arrangements (referred to as the "contract curve") that imply different divisions of the surplus created by the parties' joint profit-maximizing behavior. Any of the points on the contract curve will be characterized by the same qualitative efficiency conditions. However, many of the ILM issues which previously have been viewed as distributional are actually subsumed in an efficiency model which allows for asymmetric information, transaction costs, and sunk investments. For example, strategic behavior between workers and the firm, motivated by an interest in redistributing the ILM surplus, are treated as constraints whose costs the parties attempt to minimize by creating contractual mechanisms that reduce inefficient, rent-seeking behavior.

The paper is organized as follows: We begin by comparing the internal labor market with the standard textbook or external market. Then we discuss and attempt to integrate the contributions of specific capital, risk aversion, asymmetric information, and transaction costs. Next, we analyze contract enforcement, describe the contractual terms included in the implicit or explicit agreement to make the contract self-enforcing, address various forms of third-party contract enforcement, and discuss distributional issues. We close the paper with a summary and concluding comments.[4]

[4] The literatures we review are voluminous; our discussion is necessarily selective. We

The Internal Labor Market Compared to the External Market

Before analyzing the complexities of the internal labor market, it is useful to compare it briefly with the simple textbook model of labor markets, a model which effectively describes the labor market external to the firm. The external labor market (ELM) has two components. First, many firms participate in the external market when hiring new workers. Although these markets are segmented by the general skill of the workers (ranging from unskilled workers without a high school education to corporate administrators), they cover wide geographical regions and contain large numbers of workers and firms. In such broad markets, the potential for monopsony power by firms or monopoly power by workers is limited.[5] Second, certain industries maintain labor markets that are primarily external. The literature often cites unskilled labor markets in agriculture and retail trade as examples, but some highly skilled markets, including the construction trade and some professional occupations, are also close to the ELM norm.

In the textbook model of the labor market, firms and workers make few investments in the job or in the relationship. Hence, firms can discharge workers, and workers can quit at little cost. In the extreme case, sunk investments are zero, so the parties lose nothing by terminating their relationship.

The distinguishing characteristic of the internal labor market, on the other hand, is that firms and workers incur substantial sunk cost investments. Since these investments are not portable across firms, job immobility results. If workers were to switch jobs or firms were to discharge workers, the sunk investments would be lost. Minimizing these sunk cost losses encourages the parties to maintain their ongoing relationship.

The external labor market is the benchmark for any analysis of the ILM. It provides the opportunity costs of alternative employment for workers, and of alternative workers for firms. Workers in the ILM always have opportunities to find jobs with other firms, and these external opportunities

must, for example, omit much of the empirical literature attached to the models we review. In many cases, we reference other surveys which the interested reader may consult for more extensive bibliographies.

[5] The ELM model of perfectly competitive labor markets can be broadened to include localized monopoly power. ELM imperfections arise, not from the traditional source of few competitors, but from costly search due to imperfect worker information. The result is a distribution of equilibrium initial wages (discounted to present value over the life of the contract) rather than a single price. As long as the relevant information is too costly for the firms to provide to workers, the price differences will prevail and firms will have local monopoly power over workers. The existence of wage distributions rather than a single wage has been widely documented, beginning with Dunlop (1958).

provide limits below which their rewards cannot fall. Similarly, firms can hire new workers from the ELM and discharge workers who fail to meet work standards. Although the wages and other terms and conditions of employment are set administratively by the firm, they must ultimately rest on the opportunities for hiring new workers into port-of-entry ILM jobs from the external market. Hence, ELM economic pressures on the ILM are not repealed; they are simply rechanneled through these port-of-entry jobs.

The Nature of Internal Labor Markets

In this section, we describe four central economic factors that affect an ongoing employment relationship: (1) firm or match-specific training, (2) risk aversion, (3) asymmetric information, and (4) transaction costs. Our thesis is that all four of these factors need to be considered simultaneously in order to provide a view of the ILM that is consistent with the broad "stylized facts" of ongoing employment relations. Although all of these factors have been extensively analyzed in isolation, we believe this to be the first attempt to tie them together in an analysis of internal labor markets.

Match-specific capital. The central rationale for long-term attachments rests on firm-specific investments. Narrowly defined, these refer to investments in training that make workers more productive with their current firm than with alternative firms.[6] In the polar case, such training only increases the marginal product of workers on their current job and has a zero impact on their productivity with other firms. The result is an incentive to continue the employment relationship.

Match-specific investments is a somewhat broader category. It refers to firm-specific investments in human capital via on-the-job training, learning-by-doing, etc.; to worker-specific investments; and generally to the case in which a firm and a worker may simply have formed a "good match." This match implies a greater expected "surplus" than would result if a new random worker was inserted into the slot, or if the worker was assigned a new random job.[7]

The surplus consists of the firm's profit derived from its current

[6] These could also be "worker-specific" investments where, for instance, the employer designs programs, compensation packages, etc. to meet the needs and desires of a particular group of workers, making it costly for the workers to move to another job.

[7] Classic references include Becker (1964), Mincer (1962), and Oi (1962) for on-the-job training, and Arrow (1962) for learning by doing. Models of matching that do not explain in detail why some partnerships yield a greater surplus, but investigate the implications of the fact that the surplus can differ across worker-firm matches include Jovanovic (1979) and Mortenson (1985).

employees over and above what could be earned by recourse only to an external labor market, and the utility of the workers from the employment compensation package over and above what they could derive on the external market. The goal of the worker-firm coalition is to maximize this surplus subject to constraints imposed by technology, information, and other features of the environment.[8]

Workers enter a firm with general (i.e., portable across firms) training. However, productivity often benefits from match-specific investments, so the size of the surplus becomes a function of the return on those investments. The first investments in the match are the expenditures on hiring and screening that allocate workers to jobs in which their productivity is likely to be highest. Specific training can then be undertaken at a level which maximizes the value of the match.

A difficulty with match-specific investments is that although the ILM is disciplined *ex ante* by the usual market forces, *ex post* there is a lock-in effect due to the investments that have been sunk into the relationship. This makes the *ex post* ILM a bilateral bargaining situation. In this context, inefficient rent seeking is possible. A particular problem involves quits or discharges designed to prevent a party from recouping past investments. To encourage joint surplus maximization, rather than self-interested or counter-productive rent seeking, the ILM must design enforceable contractual arrangements to deal with such turnover.[9]

Given these turnover costs, why do we not observe contracts that simply prohibit or directly restrict such occurrences? For example, workers (firms) would sign contracts that prohibit quits (discharges), eliminating the potential for the other party's loss of its share of the investment. The existing economics literature has not satisfactorily answered this question.

The usual explanation is that laws against worker servitude make such contracts unenforceable, but this cannot be the whole answer. The only contracts that would involve indentured servitude are those that require "specific performance," meaning that the breaching party must fulfill the specific terms of the contract. Few commercial contracts are enforced in this way; instead, the breaching party pays damages. This damage remedy could also be used to compensate a breach of match-specific investments.

A second explanation rests on the fact that future events, such as changes in tastes or skill, could make fixed employment contracts inefficient; that is, not all turnover is inefficient, *ex post*. Such contingencies, however,

[8] Although distribution (i.e., how to divide the surplus between firms and workers) is important, we choose to give it less attention here.

[9] Studies on the implications of the specific-training model for turnover include Hashimoto (1981), Mortenson (1978), Parsons (1972), and Pencavel (1972).

could be handled by writing the relevant contingent-claims contract which would delineate job tenure and related employment issues as a function of future events. A difficulty with this solution is the transaction cost of writing contingent-claims contracts (a subject we address later in this paper).

The ILM's answer to turnover is to deal with it *indirectly* through wage or compensation policy. In his seminal study on human capital, Becker (1964) suggests that rent-seeking behavior (quits or discharges to gain a larger proportion of the surplus) could be reduced if both parties shared in the investment costs, with the goal of making their contract self-enforcing. For example, workers would invest in their own specific training to the extent that their current wage (w) is lower than their opportunity wage (ow) in the external market. The firm's investment is similarly measured by the difference between the worker's marginal product (mp) and w. The worker would be deterred from quitting, and the firm would be deterred from laying off the worker because such behavior would result in the loss of future returns on these investments.

Thus, a central result of the specific-training literature is a wedge between the marginal product and the wage (the firm's investment) and between the wage and the opportunity wage (the workers' investments). The wedge reflects the fact that the returns on investments occur later than the investment costs. A continuing pattern of such investments produces the familiar upward-sloping age-earnings profile. Internal promotions can similarly be explained by this investment pattern.

Although the parties primarily set the efficient level of turnover indirectly through compensation policy, recent legal innovations have tilted toward third-party enforcement for certain types of quits or discharges. For example, firms are increasingly relying on "key-employee" or "noncompetition" clauses that prevent workers from trading on their industry-specific knowledge. This most often affects managerial workers, the group that might be expected to receive the greatest amount of match-specific training (e.g., see Closius and Schaffer, 1984). Workers, on the other hand, are more actively pursuing court redress for losses due to "wrongful discharge." This trend raises broader issues which we discuss further below. (See also Krueger, 1988.)

Risk aversion. The "implicit contract" literature that began with Azariadis (1975), Baily (1974), and Gordon (1974) was not based explicitly on training or specific capital, but on risk allocation between employers and their workers. Whether due to better access to financial capital markets or simply to different attitudes toward fluctuations in income as stressed by Knight (1921), employers are assumed in these models to be typically less risk averse than workers.

Efficient risk sharing thus requires that compensation be smoothed. Smoothing means that mp will vary by more than w, and at any point in time mp need not equal w. Hence, the risk-sharing model, like the match-specific investment model, predicts divergence between mp and w.[10] This divergence has different implications in the two models in terms of the sequencing of pay. Only match-specific investments explain why wages increase with age. However, the profile could also exhibit a high variance. Indeed, absent risk aversion on the part of workers, wages may vary as much as profits.

Risk aversion converts the firm-worker partnership into a partnership in which workers effectively become a "limited partner" or a "secured creditor" whose payment is guaranteed against fluctuations in output or job performance. Hence, risk aversion must be added to match-specific investments to explain an age-earnings profile that is smoothed as well as upward sloping. Both factors are needed to account for the empirical regularities.

Still, risk aversion *in itself* cannot be the single basis for a continuing employment relationship, given that much of this insurance function could be accomplished outside of a firm's ILM. Insurance policies, including those concerning life, health, disability, and unemployment could, in principle, all be written by private carriers. Similarly, as is currently the case with social security, all retirement plans could be run by agencies outside of the firm or through spot market contracts.

Yet there may be reasons for incorporating risk sharing into the ILM once an ongoing relationship is in place. For example, using the ILM to perform parts of this function within the firm may reduce transaction and some other costs of insurance contracting, including monitoring. That is, whether or not risk sharing is a primary reason for the initial emergence of the ILM, once the ILM is in place, it is likely to be a cost-effective method for income smoothing.

There are tradeoffs, as well as complementarities, between match-specific investments and risk aversion. The deferred compensation that is used to make contracts self-enforcing conflicts with the goal of smoothing workers' income (unless deferred compensation can be perfectly insured against future exogenous events). The presence of these tradeoffs illustrates the importance of an integrated view. Additional tradeoffs between risk aversion and asymmetric information are discussed below.[11]

[10] Several implications of the divergence between w and mp in the implicit contract model are discussed in detail in Wright (1988).

[11] It is now widely accepted that contract theory based on risk aversion does not clarify inefficiencies or unemployment based on wage rigidities Unemployment results in these

Asymmetric information. A critical problem in the ILM is the presence of asymmetric information. Asymmetric information exists when it is relatively more costly for one of the parties to observe or monitor the quantity and quality of either inputs or outputs or the state of technology and demand. In the polar case, one party's information is entirely private and unverifiable by the other party.[12] Two classic examples are: (1) workers having asymmetric information advantages in determining their work effort and (2) firms having an advantage in determining the state of the product market and technology.

If both parties cannot observe work effort or product market conditions at equal cost, cost minimization suggests allocating the collection of such information to the low-cost party. Although it seems efficient to simply have that party report the results, incentive problems arise because the party with the informational advantage can use that information to achieve opportunistic aims. For a contract to be efficient, it must resolve this dilemma: It must not only assign the information gathering to the low-cost party, but also provide a mechanism which prevents the information from being used strategically. We call a contract that resolves this dilemma a self-enforcing contract or an incentive compatible contract.[13]

Contracts that control workers' strategic behavior. It is generally assumed that workers know their own work effort, while the firm can only learn about the quality of the workers' input through costly monitoring. Since workers prefer leisure to work, they have an incentive to overstate their effort if left to monitor it themselves. Modeled as a principal-agent problem, consider a worker (agent) who produces output (y) according to the function $y = f(e,x)$, where e is effort and x is a random variable. Neither x nor e is observed by the firm (principal), although we will assume that it can observe y. If e (or x) is public information, then assuming the worker is risk averse and the firm is risk neutral, the optimal contract would have the worker expend a certain efficient level of effort in return for *constant* wage w. The effect is to make w independent of y.

models because labor is assumed to be indivisible (see Rogerson, 1988). Such unemployment may or may not be "involuntary," but it is nonetheless efficient (see Rogerson and Wright, 1988)

[12] There are models which explicitly account for verification by the other party at a positive, but finite cost See, for example, Townsend (1979).

[13] Sometimes the efficient outcome which is subject to informational constraints is referred to as the "second best" result, indicating that it is only "best" given asymmetric information. Because informational constraints are, in principle, no different from the constraints imposed by the production function or any other aspect of the environment, we will simply refer to the outcome as efficient while recognizing that all economic decisions are made subject to constraints

When information is distributed asymmetrically, however, an opportunity arises for strategic behavior by the worker. The worker is able to put forth a very low level of e (assuming leisure is preferable to hard work) and claim that the resulting low level of y is due to a bad realization of x, so she/he is entitled to the same level of w. Hence, there is no incentive to supply the correct effort.

The optimal contract in this case (under certain fairly mild regularity conditions) sets w as an increasing function of output, $w = w(y)$, $w' > 0$. This provides incentives for more appropriate effort, although it also exposes the worker to uncertain income, which is a problem if she/he is risk averse. This illustrates an important tradeoff between allocating income risk and providing the correct incentives in contracts. It is the extension of this simple model that leads to the broad problem of motivating work effort through incentive pay.[14]

Another approach is a variant of the law-enforcement model first developed by Becker and Stigler (1974). Suppose the worker has the opportunity to shirk on the job. Let b denote the benefit of such cheating to the worker. Given a level of monitoring, let p be the probability of detecting him/her in the act. Then if we make the worker post a bond of size B, as long as $pB \geq b$ there will be complete compliance.[15]

We do not, however, often see workers actually post a bond. Many workers are "credit constrained" in the sense that they cannot raise the required amount, B (see Azariadis, 1988). Consequently, the bond may take more complex forms that circumvent the credit constraint. Such bonds are descriptive of a range of actual personnel practices. For example, internal promotion hierarchies, pension plans, and other deferred reward systems can be interpreted as partial solutions to the monitoring problem. Such mechanisms are not used more broadly because, although they may provide appropriate incentives for work effort, they may also conflict with the efficient allocation of income over time, based on insurance and other smoothing considerations.

The bonding involved in the work effort problem is different from deferred compensation in the specific-training model. In particular, in the specific-training model, increased skill means that pay is sequenced so that workers' $mp > w$ in later periods; in the work monitoring problem, the bond implies

[14] There is a vast literature on principal-agent models of contracting. See Hart and Holmstrom (1987) for a state-of-the-art survey with many references

[15] Harris and Raviv (1979) study the more complicated case where the monitoring technology is imperfect; see Parsons (1986) for a discussion of these and some other models. Lazear (1981) discusses how bonding mechanisms might work over time via back-loaded wages. See also Akerlof and Yellen (1986) and Holmstrom (1983).

that w > mp in later periods (these differences are discussed in Hutchens [1987] and Medoff and Abraham [1980]).

Contracts that control firms' strategic behavior. There are also models in which firms have the informational advantage, usually with respect to the state of product demand or technology. Two general models are worth discussing. Appropriate work incentives would encourage workers to work harder when product market conditions are favorable and mp is higher. But if w were constant due to income smoothing, the firm would have an incentive to misreport the product market as being favorable, hence forcing greater work effort. As above, the misreporting problem is alleviated by making compensation (as in bonuses or profit sharing) vary with work effort. Such contracts have important self-enforcing properties because the firm does not gain by misreporting product market conditions. Here again, however, income smoothing is traded off against appropriate work incentives (see Green and Kahn, 1983; Hart, 1983; and Cooper's [1987] survey).

A second problem relates to the sequencing of w and mp (mentioned above). In this case, workers (but not firms) are in the recoupment phase on their sunk investments in later years (with w > mp). By misreporting its product market conditions as unfavorable, a firm could seek to discharge workers who are recouping on their deferred compensation. One solution is to restrict the way in which a firm can adjust to changes in product market conditions. Seniority schedules are partly a response to this problem. When the firm is investing in workers, forcing it to lay off workers according to a seniority schedule means that it must accept a loss on investment in junior workers before senior workers (with w > mp) can be laid off (see Riordan and Wachter, 1982).

The asymmetric information literature leads one to expect to find complex state-contingent contracts including self-enforcing mechanisms developed to control strategic behavior. Such contracts would specify what happens in the face of potential exogenous changes in technology or in the demand for the firm's output, and hence inputs. Combined with risk aversion and match-specific investments, such contracts would also describe the parties agreed-upon tradeoffs between income smoothing and the provision of appropriate incentives for correct reporting of asymmetric information.

The problem with this prediction is that we do not observe such complex contracts, at least not in the nonunion sector, where over 80 per cent of the work force is employed. In fact, we often observe the opposite—incomplete contracts.

Transaction costs. The puzzle concerning the absence of detailed contracts is solved by one of the factors which explains why the relationship is brought inside the firm in the first place—transaction costs. If the parties inside the firm attempt to maximize the coalition's surplus, they must obviously attempt to reduce transaction costs as much as possible (or, more accurately, as much as it is efficient to do so). Since negotiating, writing, and enforcing contracts often incur high transaction costs, complex state-contingent contracts might not be joint profit maximizing.[16]

In place of this state-contingent contract, the parties could reach an understanding on general principles, but not on specifics. This agreement could be either implicit or explicit, although in most nonunion firms it is entirely implicit. In this contracting framework, the parties deal with new events by rolling over their general understanding to these new factors.[17]

Incomplete contracts might seem to worsen problems of asymmetric information. Absent detailed, state-contingent contracts, what factors prevent opportunistic behavior by either of the parties? Perhaps the most important disincentive for strategic behavior is the repeated nature of the ILM relationship. Repeated transactions are less subject to opportunism than are short-run relationships. An opportunity for gain that results in a breakdown of the relationship is not likely to be pursued if there is much surplus to be lost or significant fixed costs to be incurred in terminating or restarting the relationship. Long-term relationships sometimes can reduce opportunities to misrepresent the outcomes of stochastic events due to the application of the law of large numbers; it is simply not acceptable to report that a certain

[16] Wilhamson, Wachter, and Harris (1975) emphasize that efficiency dictates the use of incomplete contracts and that long-term relationships are designed to economize on the real resources that are required for negotiating, writing, and enforcing agreements, as well as for adapting efficiently to certain exogenous changes in the economic environment.

[17] It is useful to underline the distinction between an implicit contract and an incomplete one. Both share the distinction of being unwritten and, therefore, both save on certain transaction costs. However, in an implicit contract, a "meeting of the minds" has in fact occurred. The contract is two-sided, showing the "consideration" which may be sufficient to make the contract enforceable. Seniority provisions (in the nonunion sector) and income smoothing over the business cycle are examples of implicit contract terms Implicit contracts are not unique to internal labor markets—most commercial contracts are at least partially implicit. Moreover, the courts have no difficulty in enforcing such contracts. In an incomplete contract, there has been no meeting of the minds. Incompleteness may involve contingencies which had not arisen before or which could be construed as unforeseeable (e.g., the liability of successor firms to honor the implicit contract agreed to by a liquidated firm). A final consideration concerns explicit terms that are not meant to be enforceable, such as a current controversy involving firms' employment handbooks that purport to describe the rights of workers in the plant. Until recently, these explicit terms were not thought to be enforceable in court, and not considered part of any contract. This is changing, particularly in the area of wrongful discharge.

advantageous outcome has occurred too often.

Reputational considerations are also frequently cited as critical in restraining strategic behavior. Obviously firms are more likely than workers to acquire reputations in the external labor market. To the extent that firms engage in strategic behavior at the cost of workers, their reputation in the external market will suffer. These firms will have to pay higher wages to attract new workers or will find it more costly to continue the contract provision that requires the workers to post a bond in the form of deferred compensation.[18]

A second control over strategic behavior is the potential for retaliation by the other party. Firms can obviously discharge workers. The more difficult issue concerns redress for workers. By deciding to shirk in response to perceived unfairness, workers can prevent a firm from realizing profits generated by strategic behavior. In the extreme, workers can engage in sabotage. Using such methods, however, clearly reduces the joint profits of the parties.

Perhaps the most powerful redress available to workers is to insist that their contracts be made more explicit and more enforceable by third parties. This effectively means that the workers will become unionized. Third-party enforcement generates transaction costs that reduce the joint profits generated by an employment relationship. Moreover, when the underlying issue is one of misreporting asymmetric information, third-party enforcement would necessitate that any asymmetric information be provided to the third party.

Summary. We have identified four factors—match-specific investments, risk aversion, asymmetric information, and transaction costs—that are important in shaping the ILM. As each of the factors explains some, but not all of the observed characteristics of ILM behavior, they must be considered collectively. An important example is that risk aversion as well as specific investments are needed to explain wage patterns that are smoothed as well as increasing with tenure.

In this integrated framework, tradeoffs and conflicts between the ILM's responses to these factors become apparent. For example, risk aversion conflicts with the fact that the wage should be an increasing function of output under asymmetric information. A second example is that the desire for detailed state-contingent contracts conflicts with transaction costs. The job of the ILM is to resolve such conflicts and tradeoffs. We would expect observable ILMs to resolve these tradeoffs differently depending on the preferences of the workers and the technology of the firm. Since these are

[18] These arguments have merit, especially with respect to the effects of reputation. See Carmichael (1984) and Bull (1987).

similar to the tradeoffs that economists analyze in their study of resource allocation, the economic model can be used to illuminate the precise tradeoffs as well as to describe the choices made by particular firms and workers.

Third-Party Enforcement: The Role of the Legal System

In this section, we discuss the methods of third-party enforcement and their efficiency properties. The types of contracts we consider include: (1) contracts in the external market (ELM contracts), (2) contracts in ILM union markets, (3) contracts in ILM nonunion markets, and (4) contract terms introduced through statutory or common law.

Commercial contracts in the external labor market. An analysis of labor contracting in the external market provides a useful benchmark for understanding ILM contracts, because of the similarity of the economic relationship in the two markets. Of special interest are ELM relationships with considerable match-specific investments, such as contracts involving personal services, subcontracting, franchising, and exclusive dealerships and distributorships.

Contracts in the external market fit a prototype of detailed, state-contingent contracts predicted by the theory involving asymmetric information. These contracts include explicit formulas for dividing the surplus dependent on stochastic events, as well as enforcement methods. The parties bargain under a regime of freedom-of-contract, so that there are no mandatory standards.[19] The assumption is that the parties themselves know best what types of clauses fit their needs. Finally, these ELM contracts make considerable use of third-party enforcement.

Third-party enforcement is treated within the law governing commercial contracts. The extensive law and economics literature in this area finds that contract law is broadly consistent with economic efficiency. In cases involving contract breach, courts act to enforce the terms of the contract. Legal precedents serve as "default settings" that tell the parties how the law will interpret the contract if it is silent on the subject being contested. The default settings make it possible for the courts to enforce contracts without complicated case-by-case litigation. Similarly, standard-form contracts, which reflect industrial practice, evolve in these markets. Such "off-the-shelf" contracts are complex, but are inexpensive to use. Default settings and

[19] There are a few statutory restrictions on the freedom-of-contract. For example, in a number of states, statutory regulations may require exclusive dealerships to run for some minimal time period.

standard-form contracts reduce the transaction costs associated with contract formation.

In an ELM contract, the rules governing termination of the relationship usually are explicit. Most often the contract has a minimum duration but is terminable at will (at discrete intervals) by either party after that period. Terminating the agreement within the minimum period would involve a penalty related to the monetary value of the sunk investments. In other words, match-specific investments are protected by joint consent.

If the contract is silent on termination, the default setting is that the contract is terminable at will only after a "reasonable duration." (This is true particularly in contracts involving exclusive dealerships or franchises.) If the contract is terminated before that point, the breaching party must pay damages. Here again, the damages are frequently the value of the match-specific investments. However, an exception is made if the court finds that the breach was opportunistic, in which case penalty damages are assessed.

Detailed contracts in the union sector. As is true in ELM contracts, union contracts are relatively detailed, explicit, and state contingent.[20] Moreover, they typically provide for an arbitration process that fills in many of the gaps in the explicit contract.

The differences between ELM contracts and union contracts involve restrictions on the freedom-of-contract, including (1) a requirement that the parties bargain over "mandatory topics"; (2) a set of rules governing the use of "economic weapons" if the parties bargain to impasse over the mandatory topics; and (3) a process that allows for union certification. These standards are inalienable in that the firm cannot require as a condition of employment that workers or their union bargain away these rights. On the other hand, all of the restrictions involve process; there are no mandated outcomes. So, after bargaining to impasse on the mandatory topics, the firm can, as one of its weapons, hire permanent replacement workers who might then petition for the union to be decertified.

An interesting question is why the National Labor Relations Act (NLRA) makes substantial use of inalienable entitlements. Many traditional labor law scholars maintain that the NLRA attempts to foster industrial peace and democracy by giving workers greater bargaining rights and a guaranteed voice in the employment relationship. The assumption is that if the rights were alienable, the inequality of bargaining power in favor of the firm would

[20] Since the union sector is analyzed elsewhere, we only briefly discuss the points that are relevant to our focus on the efficiency of alternative contracting rules. For an early treatment of labor unions, see Rees (1962) For a recent general survey, see Farber (1986).

result in workers giving away those rights under duress (see Atleson, 1983).

From a purely economic perspective, it is always difficult to defend inalienable rights. The restrictions on the bargaining process would have to be based on the presumption that the government knows the efficient process and that the parties themselves would not use the same process because of some failure in the bargaining power. This argument seems strained today, but it may have made more sense during the thirties.

The efficiency argument for labor unions is that there are potential gains in having workers choose an exclusive agent-auditor to represent them in bargaining with the employer. In this context, unions lower the transaction costs by replacing worker-by-worker bargaining with a single agent. The agent also acts as an auditor who monitors the firm's use of its (asymmetric) information and reduces the potential for inefficient rent seeking. This view of unions has been stressed by Freeman and Medoff (1984).[21]

Unions and the NLRA have recently become an active topic of research in the law and economics literature. The traditional interpretation is that of the Chicago school, which argues that the intent of the NLRA was to cartelize the labor market, resulting in successful rent-seeking by workers (see Posner, 1986).

Wachter and Cohen (1988) analyze National Labor Relations Board and court decisions regarding specific contract rules. In this work, the contract terms and the NLRB and court interpretations of those terms are judged against a standard of efficiency. The conclusion is that these rules are broadly consistent with the self-enforcing contracts described above. A particularly important issue for our topic involves job tenure and the mobility of capital in the union sector. In this area, NLRB decisions leave firms with considerable unilateral freedom to determine employment levels, to relocate work across plants, and to sell assets to other firms. This is compatible with firms having an asymmetric information advantage with respect to product market conditions and technology. On the other hand, firms do not have unilateral freedom to change wage rates, a rule which is also in accord with the setting of information asymmetries. To limit strategic behavior, however, the NLRB also infers a broad obligation on the part of the firm to bargain over the effects of the decision. This mandatory bargaining would address such factors as seniority, within and across plants, and severance pay.

These legal rules are similar in substance, although not in process, to the implicit rules in nonunion contracts. In nonunion contracts, firms obviously

[21] Freeman and Medoff (1984) recognize both the positive and negative roles of unions. A recent literature suggests that the primary empirical effect of unions is successful rent-seeking in the form of high union wage premiums See, for example, Hirsch and Addison (1986), Addison and Hirsch (1989), and Linneman and Wachter (1986).

retain decision rights over total employment levels and capital mobility, but frequently accord workers protection through informal seniority provisions and relative wage stability. (See also the discussion below of the Plant Closing Act.)

In terms of process, the grievance procedure in most union contracts gives union workers greater protection with respect to termination than that found in the nonunion sector. Union workers can contest a dismissal using the grievance process, while in the nonunion sector an inference of employment-at-will means that there is no formal recourse against a dismissal. (However, see the discussion below on wrongful discharge suits under common law.)

The efficiency of NLRA contract law in the union sector has implications for the long-run market shares of unions. The NLRA left considerable room for workers to remain in nonunion firms where contract rules are very different. The result has been, and continues to be, competition between union and nonunion firms and even between union and nonunion subsidiaries of the same firm. The competition pits the legal process rules of the NLRA, the resulting detailed explicit contracts in the union sector, and any noncompetitive contractual outcomes against the nonunion contracts which we describe below.[22]

Incomplete contracts ın the nonunıon sector. As discussed above, in most internal labor markets in the U.S., contracts are largely incomplete and, where provisions exist, they are typically implicit. This fits with the predictions of the transaction costs model, which views ILMs as promoting the joint surplus through savings on contract costs.

The bipolarization of the incomplete contracts in the nonunion ILMs versus the detailed state-contingent contracts observed in ELM contracts and in the union sector is puzzling. Analyzed from the perspective of the nonunion contract form, three explanations merit attention.

First, the degree of incompleteness may reflect the underlying rationale for creating internal labor markets. As stressed above, a comparative

[22] An obvıous questıon for contract theory to explain ıs why some sectors become unionized and write very detaıled contracts while others remaın nonunıonized and wrıte almost no contracts at all Part of the answer ıs almost certainly historıcal The craft and ındustrial unions formed durıng the thirtıes were ın the more mature manufacturıng, constructıon, mınıng, and transportatıon industries. Other servıce-producıng sectors were less ımportant during that time period and largely remained nonunıonized The only sıgnificant sector that became unıonızed after the fifties was the government sector. Presumably the choıce between unionizıng or not would also reflect the ındustry-specıfic costs and benefits of unionizing. These are typıcally described as the Hicks-Marshall condıtıons (See, for example, Ehrenberg and Smıth [1988].) Unfortunately, little attempt has been made to analyze whether these condıtıons make sense ın the broader context of labor contract theory

advantage of organizing activities inside the firm is to save on the transaction costs that occur in writing explicit ELM contracts. From this perspective, it is not the incomplete contracts in the nonunion sector that require explanation but the detailed contracts in the union sector.

A second factor is the NLRA rule that makes it unlawful for employers to dominate, assist, or interfere with the formation or administration of any labor organization. If a firm in the nonunion sector were to *negotiate* a contract with its workers, it would be an unfair labor practice, and the activity would be enjoined. The outlawing of company-dominated unions reflects the opinion during the thirties that such unions only serve to thwart true collective bargaining. In today's environment, it is certainly intriguing to ask whether the nonunion sector would be "more organized" without these legal restrictions.[23]

Finally, bipolarization may arise from the difficulty of writing very partial contracts. If a contract is to be enforceable, it must be specific enough to enable the courts or the third-party arbitrator to draw guidance from the terms to apply to the area in dispute. Contracts that are largely incomplete, but contain a few enforceable clauses, are vulnerable to misguided rulings. If the parties are to write a contract, it is thus likely to be detailed.[24]

The duration of the employment relationship in the nonunion sector is based on a default setting of employment-at-will. Under this doctrine, an employer has complete freedom to terminate an employee for any reason. Recently, dents in that precedent have occurred as discharged workers have sought relief through a claim of wrongful discharge. Under wrongful discharge, a plaintiff can request reinstatement or damages if the court finds that an implied contract exists between the parties. In such cases, the court attempts to learn the terms of the unilateral contract signed by the employee upon joining the firm. In particular, it looks for evidence (sometimes from employee handbooks) that the firm appears to be offering a long-term

[23] Of course, employers can still draft explicit contracts. This can be accomplished by bargaining with individual workers over their specific terms and conditions of employment However, to devise a contract covering many workers, the agreement could not reflect bargaining between the parties. Rather, employers would unilaterally write a contract to which the workers would effectively agree by accepting the offer of employment.

[24] Contract breach often occurs over an event whose consequences were unforeseen at the time of contract formation. If the event is neither explicitly nor implicitly covered by the contract terms, the court may decide the outcome by filling in the gaps in the contract. The court attempts to determine how the parties themselves would have dealt with the event if they could have foreseen it during the contract formation. When a contract is largely incomplete, the court is less likely to be able to fill in the gap. The result is that the contract cannot be enforced by the court, or if enforced, might be prone to errors and hence inefficiencies.

contract.[25]

Laws regulating union and nonunion ILMs. Prior to the seventies, Congress was reluctant to intervene in the employment relationship, beyond its broad support of unionization through the NLRA. In part, this stance was based on the assumption that labor unions were the law's solution to employment contracting problems. Having chosen collective bargaining as the mechanism to resolve ILM difficulties, further statutory intervention in the form of explicit contract terms or outcomes was avoided. (See Gross [1974] for an extensive history of the NLRA.)

Statutory regulation of all ILMs, whether union or nonunion, began in earnest in the seventies when it was clear that the majority of U.S. workers would remain nonunion. The Occupational Safety and Health Act (OSHA), which established health and safety standards in the workplace, was the first attempt to regulate aspects of the employment relationship. In 1974, authority for establishing standards for pension plans was established under the Employee Retirement Income Security Act (ERISA). In 1988, the Worker Adjustment and Retraining Notification Act (referred to as the Plant Closing Act), provided standards that firms must fulfill before they can close or relocate a plant.

These statutory measures, which apply to both the union and nonunion markets, put in place the kinds of complex, state-contingent contract terms envisioned by contract theory. However, in a turnaround in policy toward the ILM, the parties themselves were not permitted to draft their own terms; instead, standards were set by third-party regulators.

There is a considerable literature on the efficiency aspects of OSHA (see Viscusi, 1979) and ERISA (see Ippolito, 1986). It can be shown, for example, that if the government agency knows both the workers' preferences and the firm's offer curve, it can set the optimal contract terms. Moreover, if the government knows this better than the parties themselves, such intervention would be necessary to reach the optimal contract. But it seems implausible

[25] The debate over employment-at-will should be differentiated from the issue of freedom-of-contract. Nonunion contracts, like other commercial contracts, operate under a broad mandate of freedom-of-contract. That freedom allows the parties to adopt whatever terms are mutually advantageous. In most areas of contracts, the law allows the parties to reach agreements, with the courts only serving as a mechanism for enforcing those private agreements. Currently, most nonunion contracts, with the exceptions mentioned above, do not explicitly indicate whether the courts should allow for wrongful discharge Hence, in the debate over wrongful discharge, at issue are the terms that the court should infer when the contract is silent. That debate is not of great import for the long run unless the courts were to decide that their default settings were nonwaivable by the parties.

that a regulatory agency would know individual preferences better than the individuals themselves; it is slightly less implausible that an agency would know better the risks, and hence, the true offer curve of the firm. A stronger efficiency argument for standards is that they reduce transaction costs, since the parties themselves need not deal with the issues. The standards are *minimum* standards; they are akin to default settings that can be raised, but not lowered. In this sense, they are akin to minimum wage laws, overtime pay, and other forms of protective labor legislation.

The Plant Closing Act is particularly relevant to this paper since it alters the manner in which the ILM contract may be terminated (see Ehrenberg and Jakubson, 1988; 1989). The Act's provisions comport closely with NLRB and court rules regarding plant closings in the union sector. The rule as formulated by the NLRB is that firms have a unilateral right to make the *decision* to close or relocate. However, the firm must bargain with the union over the *effects* of the closing. Under the Plant Closing Act, firms have the unilateral right to decide to close a plant, but they must give prior notice of that closing to the workers. The Act will thus lead to discussions (nonunion) or bargaining (union) over the effects of the decision on workers. In either case, the parties have recourse to certain economic weapons.

Common law and statutory law regulating ILMs reduced the bipolarization that has characterized union and nonunion contracts in the past. How far this trend continues will certainly remain an important topic in labor economics. Our argument is that such laws can usefully be analyzed using the principles of economic efficiency described above.

Distributional issues. Whereas we stress the efficiency aspects of the ILM contract, an alternative view argues that contract rules are primarily about battles over income distribution (or the surplus created by the contract). This is often combined with a related point that contractual terms are more about "fairness" than efficiency and that ILMs are frequently inefficient given the high transaction costs, worker immobility, and potential for strategic behavior. In the older neoclassical literature, these concerns would indeed appear as inefficiencies and about battles for income shares. This view of the ILM is most frequently found in the literature that developed around the initial work of Doeringer and Piore (1971).

In current modeling, however, efficiency is defined as a surplus maximization contract where the maximization includes constraints imposed by asymmetric information, transaction costs, and match-specific investments, in addition to the more traditional constraints imposed by technology and endowments. In other words, some of the presumed sources of inefficiency are now incorporated into the maximization process itself. Hence, when the ILM parties design rules to control each other's potential strategic behavior, they can be viewed as primarily acting to maximize the surplus. However, typically there is no unique, efficient ILM contract. Instead, within the ILM there is generally a set of contractual arrangements (the contract curve) that imply different divisions of the surplus.

Legal and statutory rules, as distinct from private contractual terms, are less likely to be about efficiency and more likely to be about income distribution. Legal rules can change the initial entitlements of the parties, thereby altering the final income distribution. Moreover, although competitive market discipline drives the parties toward efficient behavior, the same discipline does not necessarily drive legal rules. In fact, there is a considerable literature which argues there is a "market" for statutory rules, and that the demand for such rules is derived from rent-seeking behavior by interested parties who lobby for the legislation. Even in these cases, however, the rules cannot have *long-run* distributional effects in voluntary contracts of a repeated nature. Since each party is motivated to achieve at least the market rate of return on its investments, it will shift resources out of sectors with below-market outcomes. The only exception arises when the legal rule is both inalienable and governs the entire relevant labor market. Of course, in the ILM contracts with extensive match-specific investments, the short run can last for a long time.

As noted above in the discussion of union labor markets, for example, the Chicago school views the NLRA as primarily an attempt by unionized workers to redistribute income in their favor and away from capital and nonunion workers. Hence, the literature on bargaining outcomes has determined that most unions achieve a premium wage above the competitive market. Moreover, this premium has persisted for several decades. On the other hand, the long run may be approaching, as unions lose market share to the nonunion sectors of the labor market (see Linneman and Wachter, 1986).

Conclusion

In this paper, we have analyzed the functioning of internal labor markets, emphasizing the efficiency aspects of the implicit and explicit contracts that govern the relationship. In our model, the ILM exists because it furthers the utility and profit maximizing goals of the parties.

To summarize, firms and workers make sunk firm-specific or match-specific investments, such as certain types of worker training, that effectively lock them into an ongoing relationship. Due to workers' risk aversion, the parties agree to income smoothing, so much of the stochastic variation in the surplus is borne by the firm. At the same time, however, the existence of asymmetric information introduces some of the most complex problems that threaten cooperation in the ILM. Since strategic false reporting implies an advantage, neither party can rely on the other always to report their information truthfully. The result is to encourage the parties to adopt self-enforcing contract terms. The analysis of such terms represents an active area of economic research that can be applied to the analysis of actual ILM contracts. Finally, what is required to explain actual ILM contracts is the additional assumption that agents can more efficiently bring the inputs inside the ILM, rather than purchase them on the external market. This follows from transactional cost savings. More specifically, contracts can be made less explicit and less complete.

Contracting inside the firm poses difficult questions of enforcement. The parties themselves recognize and limit the potential for strategic behavior by agreeing to terms that have important self-enforcing properties. We have suggested that these self-enforcing contractual terms include many of the stylized features of actual ILM structures.

The alternative to using self-enforcing mechanisms is to write contracts that can be enforced by third parties. Use of third-party enforcement varies considerably across labor markets. For example, while union contracts and labor contracts in the external market (e.g., personal service or subcontracting) make substantial use of third-party enforcement, nonunion contracts are designed to be almost entirely self-enforcing.

There are important tradeoffs in dealing with the four factors and the related enforcement issues. On the one hand, in order to provide the correct incentives for joint profit maximization, contracts might involve investment cost sharing, deferred compensation, and compensation that depends on performance. However, such incentive terms might conflict with the goal of efficient risk bearing. A second tradeoff exists between the need for complex contingent claims contracts and the transaction costs of writing such contracts. The result of this tradeoff is a bimodal distribution of contract forms. In the nonunion sector, contracts are largely incomplete and almost entirely implicit. In union contracts and in external labor market contracts, contracts are largely explicit and reasonably complete.

Economic analysis of the ILM should be useful in evaluating specific firm contracts that are developed by the parties themselves. Also, by creating a benchmark of efficient contracting, economic analysis assists in determining the effects of regulation on the employment relationship and on the welfare of the parties. This highlights the nature of the factors that shape the ILM and the potential tradeoffs they create. Since economic analysis is designed to examine such tradeoffs, it can usefully be brought to bear in the study and in the operation of the ILM.