Carliss Y. Baldwin
Harvard University

Productivity and Labor Unions: An Application of the Theory of Self-enforcing Contracts*

In industries with sunk-cost technology, a union may seek to expropriate capital returns when the capital replacement cycle is long relative to the union's horizon. A self-enforcing counter-strategy is for investors to maintain inefficient plant in production in order to discourage higher wage demands. This strategy can protect capital returns, but is costly because it reduces productivity and decreases the industry's incentives to grow and to innovate. Alternative methods of contract enforcement include binding arbitration, incentives to extend workers' horizons, debt, and vertical integration. These reduce and sometimes eliminate the need to maintain inefficient capacity as a deterrent to wage demands.

In industries like steel, autos, and mining, capital investments in long-lived plant and equipment must be made before production can begin. To compensate investors for their capital, returns must be extracted over the facility's life. However, once the fixed plant and machinery are in place, it is economically advantageous to keep a facility open as long as revenues are greater than ongoing costs. Knowing this, an adversary such as a labor union may be able to expropriate the surplus that provides the return to capital. But in this case informed investors would not make the initial investment at all.

This paper identifies a self-enforcing counter-strategy available to investors. The strategy is to invest in less productive plant and equipment in order to discourage higher wage demands. The strategy protects investor returns on capital but is costly in that it raises average operating costs in the industry and reduces incentives to grow and to innovate.

* I thank Fischer Black, Susan Geller, Richard Ruback, and Michael Spence for stimulating discussions which influenced the content and presentation of this paper. Constructive comments by Armen Alchian, Robert Hayes, Michael Jensen, Lester Telser, and an anonymous referee also contributed significantly to this work. Versions of the paper were presented in workshops at Radcliffe, Harvard Business School, and M.I.T. and I thank the participants for their insights. The financial support of the Bunting Institute of Radcliffe College is gratefully acknowledged. Any mistakes are, of course, my own.

(*Journal of Business*, 1983, vol. 56, no. 2)

0021-9398/83/5602-0001$01.50

A self-enforcing contract establishes terms of trade between two parties so that neither side has incentive to revise the contract after the fact. Self-enforcing agreements are analyzed in Telser (1980); similar incentive arrangements are also found in Becker and Stigler (1974); Klein, Crawford, and Alchian (1978); and Klein and Leffler (1981).

The need for self-enforcing contracts arises in the context of bilateral exchange when one party in the transaction is vulnerable to predatory behavior by the other. For example, a consumer may be tricked into paying too much for low quality goods which the seller falsely represents as being of high quality. However, knowledge of his own vulnerability makes the consumer unwilling to pay the high quality price. As a result, the market for high quality goods disintegrates even though consumers are willing to pay more for them.

A self-enforcing contract is a possible solution to the market failure just described. Let the producer foresee a sequence of repeated sales to a particular set of consumers. If the present value of future sales outweighs the one-time gain from selling under false pretenses, then the producer has an incentive always to provide high quality goods. Consumers, understanding this incentive, may undertake to make repeated purchases from one supplier, confident that it is not in the supplier's interest to cheat.

The need for a self-enforcing agreement also arises when labor and capital are jointly used in a production process. The need arises because the less flexible factor, that is, the factor which cannot be easily reallocated to some other use, may be vulnerable to predatory demands by the more flexible factor. For example, suppose a special purpose plant is built. After construction, the plant can either be used for the purpose for which it was designed or sit idle. On the other hand, workers may have attractive intermediate options. They may be able to transfer their labor to other jobs or, if unemployed, may receive unemployment compensation. In this case workers' incentives after plant construction will be to demand a higher wage; capital's short-run incentives will be to give in. Special purpose capital is thus vulnerable to expropriation of its returns by the more flexible labor force. This fact was recognized and its implications first analyzed by Simons (1944).[1]

A self-enforcing wage agreement has the property that workers

1. Simons realized that his analysis ran counter to prevailing views and apparently consented to publication only after editorial encouragement. His article makes a forceful case for free competition in factor markets, particularly labor markets. The present paper, although narrower in focus, builds on and formalizes two of Simons's insights: first, that returns on durable capital are vulnerable to expropriation by a labor monopoly, and second, that capital vulnerability ultimately acts as a deterrent to investment and new capital formation. More recently, Klein, Crawford, and Alchian (1978) have shown that the situation Simons identified might be reversed: A labor force with job-specific human capital might be vulnerable to excessive demands on the part of highly mobile capital. Footnote 14 sketches a self-enforcing union strategy that uses seniority to control capital demands. This paper, however, concentrates on the case of capital vulnerability and particularly on actions investors can take to protect the capital returns.

would lose as much as they gained by demanding a higher wage. This paper explores how investment in inefficient plant and equipment can make the wage contract self-enforcing. Section I develops a simple model of investment in a productive asset and compares investment decisions when labor is competitively supplied or monopolized and when capital is short-lived or durable. It is argued that investments that would be undertaken in a competitive labor market may be forgone if capital returns are vulnerable to expropriation by a union. In general, capital is vulnerable if it cannot be readily withdrawn from production. In Section II it is shown that if capital is vulnerable, investors may benefit by having inefficient facilities in production. The investors' problem is formulated as a constrained maximization problem. An optimal inefficient strategy is derived and its properties examined. Section III relates the problem to others in the field of contract enforcement and considers various contracting alternatives (debt, arbitration, vertical integration) that may reduce the need to maintain inefficient facilities. Section IV presents conclusions.

I. Capital Durability, Labor Market Structure, and Investment

Investments justified with short-lived capital or in a competitive factor market may not be undertaken if the capital is durable and the supply of a factor monopolized. Consider an industry selling one good, whose price is determined competitively in world markets.[2] The world price per unit output is p. For simplicity, assume there is one domestic firm. The domestic market is of size M and the domestic firm behaves rationally and represents the interests of the shareholders of the firm.

The industry uses two factors: labor and capital. Technology is assumed to be unchanging over time. There is one technology termed "efficient" or "minimum cost" which has the lowest cost per unit output if capital costs are measured at the market rate and labor costs are measured at competitive rates for equivalent skill. The minimum cost per unit output is denoted c_o:

$$c_o = kI + w_o h_o, \tag{1}$$

where k = the annualized cost per unit capital;[3] I = capital investment required per unit output; w_o = the competitive hourly wage rate; h_o =

2. This assumption is made to simplify the analysis. Results similar to those found in Sections I and II can be derived for firms facing downward-sloping demand curves. However, the analysis is made more complicated by the fact that the quantity of the good supplied becomes endogenous. The competitive world price assumption may approximate reality in some capital-intensive industries (steel, autos, etc.).

3. I have normalized k to reflect the replacement cycle of capital. Thus, if r is an appropriate (risk-adjusted) market rate, and T is the capital's useful life,

$$k = \left(\int_0^T e^{-rt}dt\right)^{-1}.$$

As $T \to \infty$, $k \to r$.

labor hours per unit output. In addition to the minimum cost technology, there are higher cost technologies. For simplicity, inefficient technologies are assumed to have the same capital input as the efficient technology, and more labor input:

$$h_i > h_o \quad i \neq o. \tag{2}$$

All technologies have constant returns to scale. Finally, the cost advantage of the domestic firm is defined as the difference between the cost of the efficient technology and the world price:

$$a = p - c_o. \tag{3}$$

Capital used in production is termed "irreversible" or "long-lived" if its principal costs are incurred before production begins, if its useful life is long, and if it is specialized to production of the good in question. Capital is "reversible" or "short-lived" if any of these conditions are violated. I now consider the decision to invest in the industry under different combinations of capital reversibility and labor market structure.

Case 1. Competitive Labor Market, Short- or Long-lived Capital

The firm faces a horizontal demand curve at price p. If labor is competitively supplied, the efficient technology has per unit total cost c_o for production up to M, the size of the domestic market (see fig. 1). The domestic firm, acting rationally, invests in M units of capacity and supplies M units of output, and its shareholders realize a capital gain of aM/r, where r is a discount rate reflecting the risk of the industry. It is assumed that the domestic firm has no competitive advantage outside the domesic market, and thus investments above M are negative net present value (NPV) investments.

With more than one domestic firm the analysis is the same, except that competition among firms may cause the domestic price to fall

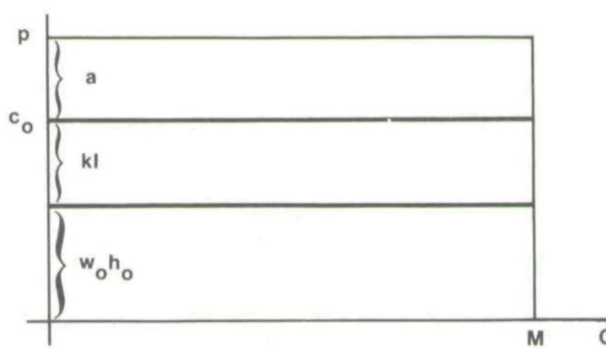

FIG. 1

below the world competitive level. With short-lived capital, the domestic price will approach c_o as the number of firms becomes large and the product market becomes more competitive.

Case 2. Monopolistic Labor Market, Short-lived Capital

I now assume that the firm needs skilled labor, whose supply is controlled by a labor union. The union is assumed to behave as a rational monopolist and seeks to maximize the total wage bill whQ where w is the hourly wage, h is labor input (hours) per unit output, and Q is units of output. The union is assumed to negotiate a uniform wage rate for all workers.

Wage maximization is an appropriate objective if workers cannot transfer their labor to another activity in the short run. They may, however, receive a fixed minimum income, for example, unemployment compensation, when unemployed. In the long run, w_o, the competitive wage rate for equivalent skill jobs, will determine the minimum rate at which labor will be supplied. Alternatively, workers might be capable of costlessly transferring their labor to another activity paying an hourly wage w_o. In this case, a risk neutral union would maximize the difference between industry wages and wages earned in the other activity. An appropriate objective function would then be $(w_I - w_o)hQ$, where w_I is the industry wage, w_o is the alternate competitive wage, and h and Q are as defined.

The form of the union's objective function does not affect the industry outcome if capital can be withdrawn instantaneously from production. In contrast, if capital is not perfectly reversible, the objective function enters the union's calculation of self interest and will affect its choice of a cooperative or a noncooperative (expropriatory) strategy. This choice is analyzed in case 3 below for different degrees of capital durability, under the imposed restriction that the efficient technology is used everywhere. Finally, the union's objective function affects the strategies available to the firm to counter expropriatory wage demands. Counterstrategies involving the use of inefficient technology are the subject of Section II.

Meanwhile, if capital is reversible and the efficient technology is used, then a competitive capital market invests in M units of capacity but the equilibrium wage in the industry is $(w_o + a/h_o)^-$.[4] In essence, the union should be able to capture the entire domestic advantage in the form of increased wages.

Even at the higher wage, M units of capacity are built because each unit of capital has a positive present value proportional to kI^+, where kI is the normalized competitive cost of the capital employed. Invest-

4. Superscripts indicate an arbitrarily small number added to (+) or subtracted from (−) the superscripted quantity, e.g., $(w_o + a/h_o)^- = w_o + a/h_o - \epsilon$.

ments up to M units of capacity thus have positive net present value, and under standard competitive capital market assumptions these investments will be undertaken.

However, if the union demands a wage higher than $(w_o + a/h_o)^-$, capital will quickly be withdrawn from the industry. Multipurpose capital will be converted to other uses, and short-lived capital will not be replaced. Withdrawal of capital is a rational response by investors to a wage demand that is too high. The union as a rational monopolist can predict this response and adjust its demands accordingly.[5]

In this oversimplified case, the withdrawal of capital is assumed to take place instantaneously and at zero cost. In cases where withdrawal of capital is costly or not instantaneous, investors will require a safety margin in the profit stream. The long-run sustainable wage is then $(w_o + a/h_o - T/h_o)^-$, where T is the normalized cost of capital withdrawal. Ironically, as long as $T < a$, less flexible capital may yield higher returns and have a higher net present value than capital that is easily reversible.

Case 3. Monopolistic Labor Market, Durable Capital

The antithesis of capital which can be withdrawn instantaneously at zero cost is special purpose, infinitely durable capital. In this case, assuming a monopolistic labor union and efficient technology ($h = h_o$), rational investors will not invest in the industry. To see this, suppose M units of capital are constructed at cost IM. By assumption, the capital is infinitely durable and valueless for other purposes. After construction, its initial cost is sunk.

When capital costs are sunk, net benefits of keeping the capacity open are positive as long as the demanded wage is less than $(p/h_o)^-$. If the union knows (or assumes) that management behaves as a profit maximizer, then it can predict that the plant will not be shut down as long as the benefits of staying open are greater than the benefits of shutting down. The union will then demand a wage $(p/h_o)^-$, just sufficient to keep variable cost less than price. In other words, the union will appropriate the return to capital. However, before building capacity, investors can calculate the union's demand and management's response. Investors can predict that a wage-maximizing union will attempt to appropriate the return to capital and that, at the time of the wage demand, capital costs will have been sunk and the profit maximizing response will be to give in. Based on these calculations, the

5. The analysis suggests an optimal replacement strategy, tied to the union contracting cycle. Investments in short-lived capital (e.g., tools and dies) should be made just after settlement of the wage contract.

investment will not compensate investors at the competitive rate and therefore will not be funded through the capital market.[6]

The outcomes of cases 2 and 3 differ essentially because of differences in the relative horizons of the union and investors. A more general version of the problem can be formulated using the calculus of self-enforcing agreements developed by Telser (1980). Define $q(t)$ as the probability that the union membership's working life extends beyond period t: $N = \int_0^\infty q(t)dt$ then corresponds to the union's expected horizon and reflects the probable duration of a cooperative agreement.[7] Let T denote the useful life of capital in place.

"Cheating" is the act of reneging on a cooperative agreement. In this model cheating occurs if, after the capital cost is sunk, workers drive the return to capital below the competitive market rate. Cheating is punished by the withdrawal of capital as soon as it is economic to do so, that is, at the end of the machinery's useful life. It is assumed that the union predicts this response and calculates benefits from cooperative and noncooperative strategies. The union will cheat if, based on its calculations, the benefits from cheating are greater than the benefits of continuing to cooperate.

Let V_t denote the union's instantaneous objective function: V is assumed to be integrable over time as well as increasing in the wage rate and the rate of total employment (hours worked). A sufficient condition for cooperative behavior (no cheating) by the union is:[8]

$$\int_0^T (V_{\sigma t} - V_{ot})q(t)dt - \int_T^\infty V_{ot}q(t)dt \leq 0, \tag{4}$$

where $V_{\sigma t}$ denotes benefits if the union cheats by expropriating some or all of the returns to capital and V_{ot} denotes benefits if the union adheres to a cooperative strategy.

Assuming that the union maximizes the present value of wages and

6. The argument that sunk-cost capital plays the role of hostage in two-party transactions has also been advanced by Klein and Leffler (1981). In their case, sunk capital provides a guarantee to the buyer against misrepresentation by the seller. Here it makes the firm vulnerable to predatory wage demands by the union.

7. A probabilistic horizon is necessary, because with a finite sequence of transactions it is almost always advantageous to cheat on the last round. Therefore, in most cases, efficient cooperation over a finite sequence cannot be self-enforcing if both parties know which transaction is the last. However, Telser (1980) shows that finite sequences for which the probability of continuing is always positive can be the basis of a self-enforcing agreement. In the case of a union, a nonprobabilistic interpretation of $q(t)$ is also possible. Suppose the union acts to maximize the sum of individual welfare for its current members. Let each member have a known, finite working life and let $q(t)$ represent the fraction of current members who will be alive t periods hence. In steady state, as long as retiring members are continuously replaced by younger members, the self-enforcing condition will be satisfied.

8. This condition is the analog of eq. (12) in Telser (1980).

predicts that it can expropriate the entire return to capital, condition (4) becomes

$$kIM \int_0^T e^{-rt} q(t)dt - (w_o h_o + a)M \int_T^\infty e^{-rt} q(t)dt \leq 0, \qquad (4a)$$

where r is an appropriate discount rate. The first term in (4a) is the present value of capital rents over the period that the plant remains in service and represents the union's gain from an expropriatory strategy. The second term is the present value of wages and surplus from the time capital wears out and represents the union's loss from the withdrawal of capital.

Short-lived capital characteristically has T small relative to the workers' horizon. For small T, a self-interested calculation of gains and losses is more likely to favor cooperation. However, as T increases, ceteris paribus, the second term in (4) or (4a), measuring the lost benefits of consistent cooperation, decreases. At some point (T^*) the gains from cheating will outweigh the benefits of continued employment: At this point the self-enforcing property of the cooperative agreement breaks down.[9] In industries with capital durability T large enough so that (4) is violated, workers' self-interest cannot be relied on to enforce nonexpropriatory behavior once the capital cost is sunk. Therefore, in the absence of some other means of enforcement, investors will not rationally commit capital to these industries on the same terms as to industries with short-lived capital where enforcement is implicit in the replacement schedule.

A counterweight to capital's durability is the union's horizon, N, which is determined by the probability of survival $q(t)$. Higher $q(t)$ (for $t > T$) signifies a higher probability that members' working lives will extend beyond the point where capital wears out. A union whose membership places a greater weight on the future is more likely to maintain a cooperative agreement with capital because of the benefits it obtains from future rounds of investment. Thus unions with younger membership are likely to be more restrained in their demands than unions whose membership expects to retire before the capital stock wears out.[10]

9. More precisely, for finite V_t, if $V_{\sigma t} - V_{ot} > 0$ all t (the incremental gain from cheating is always positive) and $N < \infty$ (the union's effective horizon is finite), then there exists $T^* < \infty$, such that for all $T > T^*$, $\int_0^T (V_{\sigma t} - V_{ot})q(t)dt - \int_T^\infty V_{ot}q(t)dt > 0$. That is, at some level of capital durability, it pays the union to cheat.

10. Simons was keenly aware of both the vulnerability of durable capital and of the impact of the horizon on a union's deliberations. According to his analysis, a union which has the "valuable hostage" of investment already sunk in durable assets, should "plan gradually to exterminate the industry by excessive labor costs, taking care only to prevent employment from contracting more rapidly than the original constituents disappeared by death or voluntary retirement" (Simons 1944, p. 8). Simons's point is that industry growth and even the replacement of capital may not be in the best interest of a union's existing membership. Using eq. (4) the interests of existing members can, to

In a stagnant or shrinking industry, younger workers have the greatest opportunity for alternate employment and are likely to be the first to leave. Unfortunately, those who remain have shorter horizons (lower $q[t]$) and therefore relatively more to gain from a noncooperative strategy. Conversely, high growth industries have both proportionately more workers in the younger cohorts and a more immediate need for investment capital. Each of these factors will tend to push a union toward a cooperative, nonexpropriatory strategy. Thus the calculus of self-enforcing agreements predicts, in the absence of other incentives or enforcement mechanisms, that unions in mature industries with durable capital will demand higher wages and benefits relative to the industry's total resources; furthermore, their demands are likely to increase as the work force ages. In contrast, high growth promotes a continual influx of younger workers; thus unions in growing industries are predicted to be more restrained in their demands and more disposed toward cooperative solutions.

For simplicity in outlining these sequences of action and response, I have suppressed numerous complicating factors that affect interactions between labor and management. In essence, I assume that the union can present a wage demand at any time and anticipate that management will respond with myopic rationality. There are no strikes. In practice, wage demands are made credible by threats of strike, although strikes involve losses to both parties. Detailed calculations of the opportunity cost of a strike to both parties would modify the analysis and result in a lower expected equilibrium wage.[11] However, the point of this section is to show that durable, special purpose capital risks having its return expropriated if a monopolist controls some other factor of production. In the absence of an enforceable contract whose term approximates the duration of capital, investments in long-term

some degree, be inferred from demographic characteristics of the union as well as technological and economic characteristics of the industry. It would be foolhardy for an investor, in estimating future returns, to ignore this calculation, for it provides information about action a union might justifiably take to further the economic welfare of its membership.

11. Except to assume that the union poses some threat that makes its wage demand believable, potential strategies involving strikes by the union or lockouts by management are not considered here. These strategies are clearly relevant and will influence any actual wage settlement. However, analyzing them raises the difficult issue of the willingness of either side to incur real costs in order to discipline the other side. To date, little is known about how limits on the cost of disciplinary actions are established: This is an important topic of research in wage bargaining and other contractual contexts. This paper oversimplifies these issues by assuming that no costly disciplinary strategy (e.g., a lockout) will be attempted on behalf of the investors. Investors, in this model, make one decision: to commit capital to an investment of some stated durability. They predict that their agents (managers) will respond as naive profit maximizers to subsequent union wage demands. Although this asumption obviously underrates the managers' potential scope of action, in reality passive investors might take a pessimistic view of their own ability to enforce a better strategy.

TABLE 1 Outcomes from Different Combinations of Capital and Labor

Labor Market	Capital Durability	Wage Rate	Profit Rate	Investment (Output)
Competitive	Nondurable ($T < T^*$)	w_o	$k + a/I$	M
Competitive	Durable ($T > T^*$)	w_o	$k + a/I$	M
Monopolistic	Nondurable ($T < T^*$)	$(w_o + a/h_o)^-$	k^+	M
Monopolistic	Durable ($T > T^*$)	0	0	0

special purpose capital may be forgone because the required competitive returns cannot be appropriated by the investors.

Table 1 summarizes the outcomes for wages and investment of combinations of competitive and monopolistic labor markets and durable and nondurable capital. In general, all outcomes except monopolistic labor market–durable capital are Pareto efficient: The opportunity is taken and the domestic advantage aM is shared between suppliers of labor and capital. In the case of a labor monopoly and durable capital, the outcome is Pareto inefficient because an opportunity with positive net present value is not undertaken.

II. A Strategy to Counter Wage Demands

In the previous section it was assumed that the industry uses the efficient technology, that is, that labor hours per unit output are h_o. In this section, the possibility of using less efficient technology ($h > h_o$) is explored. For investors facing a monopolistic labor union, it is shown that investment in less productive capacity can protect part of capital's returns by providing a deterrent to higher wage demands.[12]

When the possibility of predatory behavior exists, the vulnerable party's appropriate response depends on the objective of its opponent. A strategy which is effective against a union with no short-run employment alternatives may not be effective if workers can costlessly transfer their labor to other wage-earning activities. A union's particular objective function is thus important to the design of an appropriate counterstrategy. However, the purpose of this section is not to argue the pros and cons of any particular objective function, but is to demonstrate that for a range of objective functions, the firm can control union wage demands by retaining inefficient capacity in production. To illustrate the logic of this result, I begin by assuming the union behaves as a wage maximizer. This case demonstrates the impact of the firm's productivity choices on wage demands and also shows how the firm's "optimal" deterrent strategy can be determined. I then consider the

12. This result does not imply that investors will always or necessarily use inefficient capital in production. See the discussion of alternative contractual arrangements in Sec. III below.

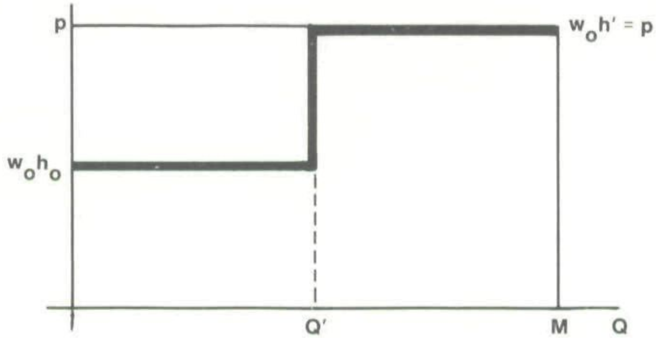

FIG. 2

problem when workers have costless access to alternative employment and indicate the effect this has on the firm's appropriate response.

The counterstrategy that uses productivity to control wage demands is motivated by a graphical illustration. Consider the industry of Section I with a monopolistic labor union and durable capital. Assume the union seeks to maximize total wages. Consider a strategy which varies the productivity of capital (h) as follows:

$$h = \begin{cases} h_o & \text{for } 0 \leq Q < Q' \\ p/w_o & \text{for } Q' \leq Q \leq M, \end{cases} \tag{5}$$

where Q' is determined by

$$p(M - Q') - (p - w_o h_o)Q' = 0. \tag{6}$$

This strategy has the following properties: (1) the wage rate demanded is the competitive wage rate w_o; and (2) although capital costs are sunk, a return stream $(p - w_o h_o)Q'$ is appropriable by the investors. Therefore if $(p - w_o h_o)Q' > kIM$, the investment in durable capital will be undertaken.

Figure 2 illustrates this policy. The dark horizontal lines indicate the marginal unit costs of output (excluding sunk capital cost). Output from zero to Q' units is produced efficiently, that is, at a marginal cost $w_o h_o$. All other output is produced inefficiently, so that its marginal cost equals price p.

If h and Q' are set to satisfy (6), a wage-maximizing labor union has no incentive to demand a wage higher than w_o. Figure 3 illustrates the impact of a wage demand $w > w_o$. With higher wages, the marginal unit cost of all output shifts upward, and costs at inefficient facilities exceed price. Management's optimal action is then to shut down the inefficient plant. The loss to total wages from shutting down $M - Q'$ units of capacity is $p(M - Q')$. From (6), even if the union were to demand and receive the maximum wage rate $(p/h_o)^-$, gains from higher wages (area G in fig. 3) would fail to offset the losses from the reduction in industry

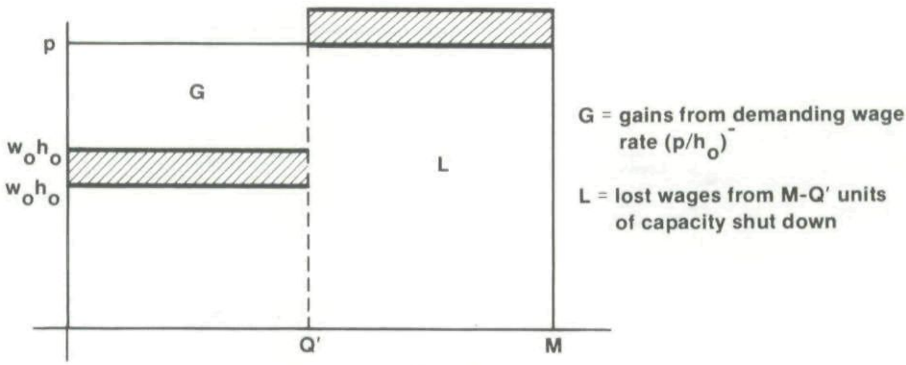

FIG. 3

capacity (area L in fig. 3). Anticipating this, a wage-maximizing union would not demand the higher wage.

The argument shows that for investors confronting a monopolistic labor union, a strategy using both efficient and inefficient technologies may dominate using the efficient technology alone. However, the strategy depicted in figure 3 is only one of a large number of possible combinations of efficient and inefficient capacity that will hold wage demands at the competitive rate. Among all strategies combining efficient and inefficient capacity, there exist some which maximize the net present value of the investment. An optimal productivity schedule is defined as a functional relationship between unit labor hours h and output Q that maximizes the value of capital subject to the constraint that the labor union has no incentive to demand a higher wage.

To derive the investors' optimal productivity schedule, the problem is formulated as a constrained optimization problem. For a wage-maximizing union, the investors' problem is:

$$\max_{h(Q)} \quad pM - w_o \int_0^M h(Q)dQ, \tag{7}$$

subject to:

$$h_o \leq h(Q) \qquad \forall Q \in [0, M] \tag{7a}$$

$$w_o \int_0^M h(Q)dQ \geq w_x \int_X h(Q)dQ \qquad \forall X \subset [0, M] \tag{7b}$$

$$p - w_x h(Q) \geq 0 \qquad \forall Q \in X \tag{7c}$$

In (7), investors maximize contribution (price less the cost of labor), which is assumed to be constant over time and is therefore proportional to the value of the firm. Industry capacity is represented by the interval $[0, M]$ and X is defined as the subset of total capacity which remains open on a wage demand w_x. If more than one wage leads to the subset X remaining open, w_x is the highest such wage.

Constraint (7a) is technological and requires labor input per unit output to be greater than or equal to labor input under the efficient technology. Constraint (7b) reflects the fact that the union can demand a higher wage ($w_x > w_o$) but that some capacity may then be shut down. From the investors' perspective, the optimal productivity schedule must have the property that the implicit wage contract is self-enforcing; that is, the union has no incentive to demand a higher wage. The union is assumed to maximize the total wage bill; thus total wages given wage rate w_o and capacity M must be greater than or equal to total wages under every alternative configuration of the wage rate and capacity. Constraint (7c) defines management's shutdown rule. For a given wage demand w_x, all capacity with variable cost $w_x h(Q)$ higher than price is uneconomic to operate and will be shut down. Units of capacity which remain open (the set X) are those whose unit variable costs are less than or equal to p.

Given any schedule, units of capacity may be reordered according to their productivity, from most to least productive. Most productive corresponds to fewest hours worked and conversely. Therefore, without loss of generality, $h(Q)$ can be restricted to the class of monotonic, increasing functions defined on the closed, bounded interval $[0, M]$. Within this class, proposition 1 establishes an inverse relationship between wage demands and productivity:

PROPOSITION 1: Define x as the least efficient unit(s) of capacity in the subset X. Because h is monotonic, the set X can be identified with the interval $[0, x]$. Then, (*a*) $w_x = p/h(x)$; and (*b*) $h(M) = p/w_o > h_o$.

Proof.

a) If $w_x > p/h(x)$, then if wage w_x is demanded, by management's rule (7c), the xth unit of capacity will be shut down. In that case, x cannot be an element of X. If $w_x < p/h(x)$, the union may demand $w' = p/h(x) > w_x$, and no capacity will be shut down. This contradicts the definition of w_x. Therefore, $w_x h(x) = p$.

b) Note that w_o must be the wage demand associated with keeping all capacity open. The inequality is implied by the fact that, with efficient capacity and competitive wages, the industry has positive profits. Therefore $w_o h_o < p$.

Proposition 1 shows that at the margin, higher productivity (lower $h[x]$) leads to higher wage demands and lower productivity to lower wage demands. In order to hold wage demands down, the optimal schedule must include some inefficient capacity. This result does not depend on the union's particular objective function, but holds in all cases where both wages and employment (hours worked) have positive value to the union.[13]

13. Formally, let the union's objective function be $V(H, w)$, a function of total employment hours (H) and the wage rate (w). Proposition 1 holds as long as $\partial V/\partial H > 0$ and $\partial V/\partial w > 0$.

Proposition 1 implies there will be both efficient and inefficient capacity in production. Total capacity $[0, M]$ can then be partitioned into efficient capacity, identified with the subinterval $[0, Q^*]$, and inefficient capacity, identified with $[Q^*, M]$. Because Q^* can be any point in $[0, M]$, there is no loss of generality from the partition. Let W now denote total wages. For a given Q^* and schedule $h(Q)$ defined on (Q^*, M), the wage bill can be written

$$W = w_o h_o Q^* + w_o \int_{Q^*}^{M} h(Q)\, dQ.$$

Because the sum of profits and wages is a constant, maximizing profit is equivalent to minimizing wages. This leads to:

PROPOSITION 2: For a wage-maximizing union the optimal schedule (defined by Q^* and $h[Q]$) satisfies (*a*) $pQ^* = W$, and (*b*) Profits $= p(M - Q^*)$.

Proof.

a) Let $x = Q^*$. Constraint (7b) implies

$$W - w_{Q^*} h_o Q^* \geq 0.$$

By proposition 1, $w_{Q^*} = p/h_o$, thus $W \geq pQ^*$. But, if W is strictly greater than pQ^*, Q^* can be increased without violating the constraint. This leads unambiguously to a reduction of wages and an increase in profits. Therefore $W = pQ^*$.

b) This proof follows from (*a*) because Profits $= pM - W$.

Figure 4 illustrates proposition 2. If the union demands wage p/h_o, then only efficient capacity $[0, Q^*]$ remains open. The union gains the difference $(p - w_o h_o)Q^*$ (denoted by area G), but loses wages from units shut down $w_o \int_{Q^*}^{M} h(Q)dQ$ (denoted by area L). To provide the disincentive for the higher wage demand, the gain must be less than or equal to the loss. However, if G is strictly less than L, it is always possible to improve the policy by allocating more capacity to efficient

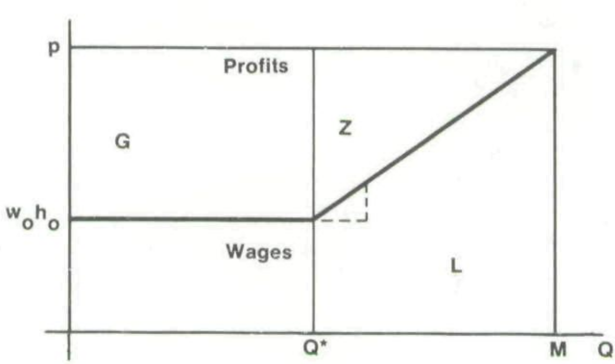

FIG. 4

production. Such an improvement is depicted by the dotted triangle in figure 4: It leads to higher G, lower L, lower wages, and higher profits. An optimal policy cannot be improved on and thus, at a minimum, must have $L = G$. Profits are equal to the area $G + Z$; given $L = G$, they are also equal to $L + Z$ or $p(M - Q^*)$.

Proposition 3 characterizes the firm's optimal productivity schedule further.

PROPOSITION 3: Consider a schedule, given by the point Q^* and a function h defined on $[Q^*, M]$, with the following properties:

$$h(Q^*) = h_o, \tag{8}$$

and

$$w_o h_o Q^* + w_o \int_{Q^*}^{M} h(Q)dQ = \frac{p}{h(x)}\left[h_o Q^* + \int_{Q^*}^{x} h(Q)dQ\right] \tag{8a}$$

for all $x \in [Q^*, M]$.

a) This schedule cannot be improved on within the constraints.
b) The following satisfies (8) and is therefore an optimal schedule:

$$\begin{aligned} Q^* &= M/[1 - \ln(w_o h_o/p)], \\ h(Q) &= \frac{p}{w_o} e^{-(M-Q)/Q^*}, \ Q^* \leq Q \leq M. \end{aligned} \tag{9}$$

c) The optimal schedule is not unique.

A productivity schedule satisfying (9) makes the wage bill invariant to the wage demanded. In other words, for every wage demand $w_x = p/h(x)$, labor's gain from higher wages is always just offset by the loss from shut down capacity. The proof that such a schedule is optimal is found in Appendix A. The essence of the proof is to show that deviations from the proposed schedule cannot decrease wages without violating a constraint.

Figure 5 graphs the firm's marginal cost curve (WW) under the optimal policy. The implied industry wage rate is w_o. The figure also shows a translated marginal cost curve (TT) corresponding to a wave demand $w_{TT} > w_o$. Under the new cost structure, $M - x$ units of capacity are uneconomic to operate $[w_x h(x) > p]$ and therefore will be shut down. A property of the optimal schedule is that the union's gain from increased wages is always just offset by lost wages from shut down capacity ($G = L$). Thus the union has no incentive to demand a higher wage.

Propositions 1–3 show that if the supply of labor is controlled by a wage-maximizing union then the optimal capacity configuration from the investor's standpoint may include inefficient capacity. The purpose of the inefficient plant is to provide a disincentive to higher wage de-

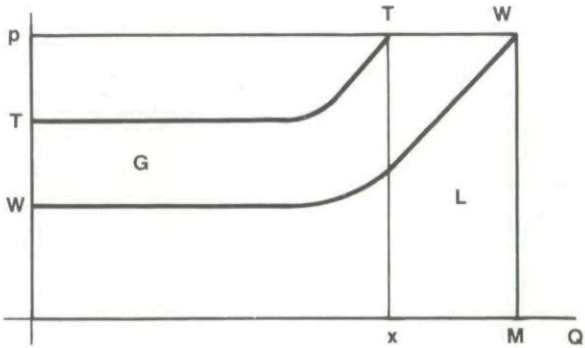

FIG. 5

mands by the union. In the absence of other means of enforcement, the disincentive is a necessary prerequisite to investment in durable special purpose capacity. Without the disincentive, after the capital is placed in operation it would be in the union's interest to appropriate most of the capital returns. A strategy such as (9) protects a portion of the profits of the industry and thereby provides a guarantee of a minimum return to capital.[14]

The strategy of deterrence through inefficiency is also useful when capital is short-lived. There, it represents a unilateral strategy that the suppliers of capital can fall back on; that is, it provides investors with a credible threat. A threat does not need to be implemented if both parties recognize its validity; thus one would not expect to see inefficient, short-lived capital actually in use. On the other hand, the presence of the threat implies that the union may not be able to appropriate the entire surplus aM (as the analysis of Sec. I indicated). Instead, the surplus would be shared between labor and capital; profits under the optimal inefficient strategy would determine the investors' minimum share of the surplus.

Value-maximizing Union

In this subsection I consider how the existence of alternate employment opportunities changes the union's objective function, and how

14. Conversely, when capital is short-lived but workers must invest in job specific human capital (must acquire a skill), all the arguments given in the text may be reversed, as workers become vulnerable to expropriation of their human capital by the providers of physical capital (see Klein et al. 1978, p. 313–15). A unilateral strategy a union might employ is to set up an implicit marginal supply function, so that if capital demands higher returns (a lower wage) the most productive portion of the work force leaves. The implicit contract would be self-enforcing if lost profits from decreases in the workforce were always large enough to offset potential gains from across the board wage reductions. This mechanism may partly explain the purpose of seniority systems. (It is important to note that the "seniority strategy" does not require actual inefficiency in production, and for that reason it is less costly to the industry and society than the strategy of deterrence through inefficiency of capital in place.)

such a change affects the counterstrategies available to the firm. Proposition 1, which established an inverse relationship between productivity and wage demands, did not depend on the particular objective function assumed. Thus, given a different union objective function the firm will still maintain inefficient capacity as a disincentive to excessive wage demands. However, if workers can costlessly transfer their labor, the firm will need to set an industry wage rate higher than the competitive wage w_o.

Assume that workers face two labor markets. In one, the union exercises a monopoly over the supply of labor; in the other, a competitive hourly wage w_o prevails. Workers can costlessly transfer their labor from one job to another. When alternate employment opportunities exist, pure wage maximization is not an appropriate objective of the union. A more reasonable objective is to maximize the excess earned by working in the industry over what the workers could earn elsewhere. This objective is similar to that of a firm which maximizes the present value of investments net of the opportunity cost of capital, which suggests the term "value maximization" for the new objective function. Formally, I define value maximization: max $(w_I - w_o)hQ$, where w_I is the industry wage rate, w_o is the alternate wage rate, h is average labor input per unit output, and Q is industry output. Given this union objective function, the firm's constrained optimization problem is like (7), except that w_I appears in the profit function, and constraint (7b) is replaced by

$$(w_I - w_o)\left[h_oQ^* + \int_{Q^*}^{M} h(Q)dQ\right] \geq \left[\frac{p}{h(x)} - w_o\right]\left[h_oQ^* + \int_{Q^*}^{x} h(Q)dQ\right]. \tag{10}$$

Analysis of the new optimization problem is complicated by the fact that the firm now has three decision variables: Q^*, $h(Q)$, and the wage rate w_I. There is, in fact, no closed-form representation of an optimal policy. However, it is straightforward to show that the optimal policy derived for the wage-maximizing union is not effective for a value-maximizing union and that the new optimal policy involves an industry wage rate significantly higher than the competitive wage w_o. (In particular cases the optimal policy may be solved numerically using dynamic programming or optimal control methods.)

To see that the optimal policy given a wage-maximizing union is not effective given a value-maximizing union, refer to figure 5. Suppose the union demands wages w_{TT} yielding the translated marginal cost schedule TT. As before, the firm shuts down $M - x$ units of capacity, but the unemployed workers can now fall back on jobs outside the industry paying the same wages. Thus instead of lost wages corresponding to area L, the workers experience no loss whatsoever. Since the gain

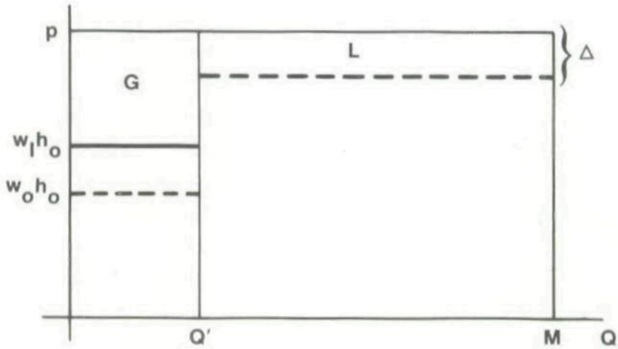

FIG. 6

from a higher wage within the industry (area G) is strictly positive, the union will always have incentive to demand a higher wage.

The question then arises, Is there any capacity configuration which supports the minimum wage rate w_o? In fact, there is. Consider, for a moment, discontinuous strategies of the form

$$h = \begin{cases} h_o & \text{if } Q \leqslant Q' \\ p/w_I & \text{if } Q \geqslant Q' \end{cases} \tag{11}$$

Figure 6 illustrates this strategy for an arbitrary w_I. Solid lines indicate the strategy; dotted lines indicate earnings at the alternate wage w_o. If the union demands a wage higher than w_I, $M - Q'$ units of capacity will be shut down. The most the union would gain by such a demand is total industry profits (area G); this is offset by a loss of excess wages on the facilities shut down, equal to area L. For the schedule to be an effective disincentive, the union's gain must not exceed the loss; that is, the firm must set $G = L$.

Given any wage rate w_I, a discontinuous schedule can be constructed to provide the necessary disincentive. However, for low wage rates the vertical distance $\Delta = p - p\ w_o/w_I$ becomes small and consequently area L approaches zero. To maintain the disincentive, the profits of the firm (G) must also approach zero. Maximum profit, subject to $G = L$, corresponds to a wage rate significantly higher than w_o. For the class of strategies given by (11), maximum profits are attained when

$$w_I = \sqrt{pw_o/h_o}$$

$$Q' = M/2.$$

This analysis carries over to more general strategies: Although there is always some strategy that supports a wage rate w_I, very low wage rates require firm profits also to be small. Optimal policies will involve an industry wage rate significantly above the minimum w_o. (It can be shown that while there is no unique optimal policy, there is a unique optimal industry wage: Multiple optimal policies with different wage rates do not arise.)

The Cost of Inefficiency

In an industry where the supply of labor is controlled by a union, investors can protect their returns by maintaining inefficient as well as efficient capacity in production. This result applies both when workers have access to alternate employment and when they do not. It also does not depend on the particular form of demand curve assumed. Similar results can be obtained for downward sloping demand curves where both the quantity of labor supplied and total output are endogenous.

Inefficient plant thus functions as a deterrent to excessive wage demands. In the absence of enforceable contracts, some form of deterrent is necessary to get rational investors to invest in durable capital. However, inefficiency imposes an opportunity loss on the industry. The loss is measured by the difference in the surplus under the constrained and unconstrained productivity schedules. Thus, for a wage-maximizing union:

$$\text{Loss} = \text{Unconstrained Surplus} - \text{Constrained Surplus}$$
$$= aM - [p(M - Q^*) - kIM].$$

Using the fact that $a = p - kI - w_o h_o$, the opportunity loss is proportional to the difference between total wages paid to the union and competitive wages: Loss $= pQ^* - w_o h_o M$ where Q^* equals $M/[1 - \ln(w_o h_o/p)]$. Considered in relation to total industry revenue (pM), the magnitude of the opportunity loss is determined by the ratio of competitive labor input to the price of output ($w_o h_o/p$): Lower values of $w_o h_o/p$ imply higher fractions of revenue lost. For values of $w_o h_o/p$ ranging from 25% to 75%, the opportunity loss ranges from 17% to 3% of total revenue.

In a dynamic context, the need to maintain inefficient capacity also results in reduced incentives to grow or to invest in technologies which improve labor productivity. However, the industry will have increased incentives to develop labor-intensive technologies, because any technology with a high marginal cost of output can provide an effective deterrent.

Maintaining high marginal cost capacity can thus impose substantial opportunity costs on an industry in the form of higher average costs and reduced incentives to expand capacity or to invest in laborsaving technology. These opportunity costs are a significant incentive to design an efficient contractual alternative to the inefficient strategy. The following section discusses various contractual methods of reducing the productivity loss. Other authors have treated related issues of contracting with great clarity and depth; my discussion is therefore brief and necessarily draws heavily on the foregoing research.

III. Contractual Solutions

Jensen and Meckling (1976) assert that a firm (and by extension, an industry) is a "legal fiction which serves as a nexus for contracting relationships" between owners of capital, labor, material inputs and consumers of output. This paper has sought to isolate one particular relationship—that between a labor union and investors—in order to examine the enforcement mechanisms that would lead to a stable contractual relationship over time. It was argued that, in the case of durable capital, the power of investors to enforce a fair contract is severely limited. If the capital markets are characterized by rational expectations, this in turn causes otherwise profitable investment opportunities to be forgone.

Fundamentally, the problem is one of imperfectly specified property rights. In negotiations with a union, investors can exercise ownership rights over physical property (i.e., they can shut down facilities), but they have no indisputable claim to a stream of future returns. This situation forms an interesting contrast with that of a regulated industry, where the right of investors to receive a fair rate of return is recognized and sanctioned by law.[15] To date no precedents exist which would lead to the overturning of a union contract because it unfairly expropriated returns to capital.

There are multiple ways to organize transactions to deal with a problem of ill-defined property rights and the opportunities for predatory behavior which ensue. Klein and Leffler (1981) divide these methods into three categories: (1) explicit contractual specification with third party enforcement; (2) direct (two-party) enforcement of implicit contracts; and (3) one-party organization or vertical integration. Each of these alternatives is discussed in turn.

Third Party Enforcement

Examples of third party enforcement are, first, explicit long-term contracts with the force of law and, second, agreements to submit differences to binding arbitration by a disinterested third party. In theory,

15. The major precedent in case law is the landmark *Hope* decision, which states: ". . . [T]he return to the equity owner should be commensurate with returns on investments in other enterprises having corresponding risks. That return, moreover, should be sufficient to assure confidence in the financial integrity of the enterprise, so as to maintain its credit and attract capital" (*Federal Power Commission* v. *Hope Natural Gas Company* 1944). No such protection is afforded investors in contract negotiations with a union.

the problem of capital return appropriation by a union can be remedied by an ex ante contractual agreement which fixes wages or provides for arbitration of subsequent disputes. In fact such contractual solutions may introduce costly rigidities into the industry's operations.[16]

A second problem with this alternative is that both parties must voluntarily enter into a long-term contract or consent to binding arbitration. This implies both parties must gain from the contractual agreement. In some cases it may not be in one of the party's interests to cooperate in the short run. For example, suppose an industry starts with long-lived, mostly efficient capacity. Threats of nonrenewal of the capital stock would not then pose a serious problem for the union for some time. In the short run, investors could do little to induce the union to accept a long-term contract or agree to binding arbitration.

Over time, investors can fall back on the noncooperative, inefficient strategy described in Section II. In the long run the industry might then shrink (because NPVs under the inefficient strategy are negative and no reinvestment takes place) or might slowly approach the "optimal" inefficient capacity configuration.[17] Then and only then would it make sense for the union to consent to long-term contracts or binding arbitration.

The history of the steel industry from 1959 to 1980 may be read as evidence that the process of rationalizing contractual arrangements can take a long time and can result in considerable deterioration of the industry's capital stock and competitive position. Briefly, in 1959, the steel industry sustained a 116 day strike. Subsequent wage settlements raised steelworkers' pay from approximately 127% of the national average for manufacturing workers to 164% in 1977 (Council on Wage and Price Stability 1977, p. 30). Growth in the domestic industry almost halted, while net imports rose from -1.4% of domestic consumption in 1958 to 12.0% in 1978.[18] At the same time, investment policies of steel firms came under criticism. Principally, the domestic industry was accused of not investing in new technology or in the modernization of facilities.[19]

16. See Klein et al. (1978, p. 316–18) for a discussion of the costs of contractual rigidity.

17. It is important to recognize that investors may "back into" the strategy of deterrence through inefficiency, simply by failing to replace or maintain the capital stock. On the other hand, if consciously employed, the strategy explains such anomalies as adding new, efficient capacity while at the same time neglecting profitable opportunities to upgrade existing facilities. Even if investments to upgrade existing plant and equipment have high NPVs, the firm may refuse them in order to maintain inefficient plant as a deterrent.

18. Before 1959, the United States was a net exporter of steel.

19. But in fact the U.S. steel industry made over $30 billion in capital expenditures between 1960 and 1978, most of it to modernize existing facilities. A more detailed comparison of U.S. investment with Japan's reveals some interesting patterns: U.S. industry lagged in installing basic oxygen furnaces (BOFs) and in adopting continuous

In 1974, when the difficulties of the domestic steel industry had become apparent, the 10 largest steel producers and the United Steelworkers made an Experimental Negotiating Agreement (ENA). The agreement provides that unsettled disputes from contract negotiations will be submitted to binding arbitration. The agreement also contains a "no-strike" clause. More recently, investment policies of the industry appear to have shifted. Increasingly, marginal facilities are being shut down, or in some cases sold to workers. However, little new investment has taken place, and the industry's main response to foreign competition has been to lobby for tariffs and import restrictions. It will be difficult, if not impossible, for the steel industry to regain the markets and technological advantages which were dissipated in the 1960s.

A somewhat different form of third party enforcement involves the use of debt. This alternative has the advantage of being unilateral; that is, management can generally change a firm's capital structure without the consent of the union. For debt to play a role in enforcing a wage contract, workers must perceive real costs (such as interruptions in employment) to be associated with the firm's failure to meet financial obligations. As an extreme example, suppose workers believe that in the event of default or bankruptcy, all capacity would be shut down. In this case, if debt is substituted for equity in the example of Sections I and II, a portion of the return on equity (kIM) is converted into an ongoing cost and the marginal cost of output increases across the board. A higher marginal cost, all other things equal, implies that a firm can (in fact must) produce more efficiently than before. Thus, if some bankruptcy costs are borne by workers, debt financing can reduce the loss from having to maintain inefficient capacity. However, the strategy is self-limiting, because real bankruptcy costs imply an optimal level of debt which is less than 100% of the firm's capital. Still, judicious use of debt may be an effective way for investors partially to control wage demands.

casting technology (despite very high prospective rates of return on this technology at some facilities). Continuous casting and BOFs both involve large capital outlays "up front" and are very durable. The United States lagged behind Japan also in shutting down inefficient open-hearth furnaces. By 1978, a "staggering" 30%–40% disparity existed between the costs of steel produced at the best and worst U.S. facilities. At that time, it was estimated that at least 20% of all U.S. capacity would be retired over the following 5 years because it was too costly to operate or to bring into compliance with environmental regulation. In contrast, the U.S. industry was quick to adopt less capital-intensive, scrap-based, electric furnace technology. Much of this investment took place at independent "mini-mill" facilities. In general, mini-mills are owned by independent producers and are not parties to the industry-wide collective bargaining agreement. Frequently they are not even unionized. Scrap-based mini-mills began to operate on a significant scale in the 1960s and during the sixties and seventies were the high-growth, high-profit sector of the U.S. industry. Of course, conflicts between labor and capital were not the only factors affecting steel industry investment patterns in these decades. The allegedly hostile attitude of U.S. government was another cause for concern in evaluating prospective investments in new capacity and may have also served to deter investment in durable "hostage capital."

Two-Party Enforcement of Implicit Contracts

The stopping of repeated purchases can enforce implicit contracts between two parties.[20] However, for this method to work, the party which guarantees the implicit contract (e.g., a firm which guarantees the quality of goods it sells), must have demonstrably less to gain from cheating than from adhering to the contract. The present value of profits on repeated purchases must be greater than the one-time gain from selling under false pretenses and forgoing all future sales.

In a union-dominated labor market, contract enforcement by stopping repeated purchase is effective when capital is nonspecialized or short-lived. As shown in Section I, when capital is readily reversible, an excessive wage demand leads to withdrawal of capital from the industry, effectively terminating employment. In considering a wage demand, a union can make the same calculation as a firm that represents its goods as being of high quality; that is, it balances a short-term gain in wages against the benefits of continuous employment. If the initial period of employment (replacement cycle) is short, the benefit of restraint outweighs the gain from a predatory wage demand, and the employment contract will be self-enforcing. However, as the replacement cycle lengthens, the workers' calculation of relative gain will be driven to favor the opportunistic strategy.

The productivity strategy developed in Section II is an effective enforcement method when capital is specialized and the replacement cycle is long. In this case, enforcement is again based on nonrepeated purchase, but, instead of being tied to capital's replacement cycle, termination of employment is related to the marginal cost (productivity) of capital units in place. Variable productivity makes partial shutdowns of capacity a credible response to excessive union wage demands. This makes the threat of termination once again significant in the union's calculations. On the other hand, as was shown in Section II, the strategy is costly: It reduces the efficiency of the industry and curtails incentives to invest in new capacity and cost-reducing R & D.

Although inefficient, the productivity strategy is unilateral and effectively controls wage demands independent of the labor union's horizon. In fact only noncooperative, potentially inefficient strategies can be self-enforcing over all possible horizons.[21] However, in Section I it was argued that expropriatory behavior is in a union's best interest only if the union's horizon (members' working life) is short relative to the capital replacement cycle (see eq. [4]). This suggests that lengthen-

20. The role of repeat purchases in ensuring quality is identified in Heal (1976). The importance of repeated transactions to the enforcement of two-party contracts is further discussed in Telser (1980) and in Klein and Leffler (1981). A similar incentive mechanism is presented in Becker and Stigler (1974) as a means of ensuring against malfeasance by an employer.

21. See Telser (1982, p. 2.4).

ing the union's horizon might lead to a self-enforcing agreement that also permits an efficient allocation of resources.

A union's horizon may be extended by giving current workers an interest in the industry's performance after they retire. One way to accomplish this is to allow workers to sell their jobs.[22] For this alternative to be effective, the industry wage rate w_I must be higher than the competitive wage rate w_o, so that the right to work is an asset with positive value. To illustrate, assume the union is a risk neutral value maximizer, that it predicts it can expropriate all the profits of the industry, and that efficient capital is used everywhere ($h = h_o$). In the absence of special incentives, the union's calculation of gains and losses is[23]

$$(p - w_o h_o)M \int_0^T e^{-rt}q(t)dt - (w_I - w_o)h_o M \int_0^\infty e^{-rt}q(t)dt \gtreqless 0. \quad (12)$$

If (12) is greater than zero, it pays the union to expropriate current returns to capital and forgo benefits of future reinvestment. As before, the union's horizon is determined by the survival rates $q(t)$. Lower $q(t)$ means that future benefits are less valuable (because the membership may not be there to claim them), and the union therefore has greater incentives to behave in an expropriatory fashion. However, suppose each worker on retiring can sell his job. Every worker would then look forward to (1) a surplus wage stream $(w_I - w_o)$ over his working life and (2) a lump sum payment equal to the present value of surplus wages upon retirement.[24] Aggregating over all members, the union's calculation of gains and losses under the new plan is

$$(p - w_o h_o)\frac{M}{r}\int_0^T e^{-rt}q(t)dt - (w_I - w_o)\frac{h_o M}{r} \gtreqless 0. \quad (13)$$

The first term, representing the gains from expropriation, is the same in (12) and (13), while the second term, representing the benefits of cooperation, is greater in (13) than in (12). Giving workers the right to sell their jobs thus increases their incentives to cooperate. Alternatively, if w_I is initially set to make (12) nonpositive, then vesting workers with the right to sell their jobs permits a lower industry wage. Furthermore,

22. A referee suggested this method of lengthening the union's horizon.

23. To facilitate comparison, eq. (12) is written in a form slightly different from eq. (4). The first term represents total (not incremental) benefits of an expropriatory strategy from 0 to T, and the second term represents total benefits of cooperation over the whole future.

24. For a worker retiring at t the sum of (1) and (2) is

$$\frac{h_o M}{N}\int_0^t (w_I - w_o)e^{-rs}ds + \frac{h_o M}{N}\int_t^\infty (w_I - w_o)e^{-rs}ds = \left(\frac{w_I - w_o}{r}\right)\left(\frac{h_o M}{N}\right),$$

where N is the number of workers. In aggregating, recall that $q(t)$ corresponds to the fraction of members surviving to time t.

shorter horizons (lower $q[t]$) now have only the effect of reducing the gains from an expropriatory strategy. This implies that the oldest workers will now have the least, instead of the most, to gain from a noncooperative strategy.

Vertical Integration

The final method of organizing transactions to reduce expropriatory behavior is vertical integration. Vertical integration alters the payoff structure so that more costs and benefits are shared in common between transactors (in this case labor and management).[25] There is at present no market that would allow a firm to purchase shares in a labor union. However, there are no obstacles to workers purchasing equity in their firms—in fact, recently such transactions have become fairly common. Therefore, I first consider the case of workers purchasing a fraction of the firm's shares.

If the union seeks to maximize wages, but workers hold claims to a percent α of the firm's profits, then the wage incentive constraint (7b) becomes

$$w_o H(M) + \alpha\Pi\,[w_o, H(M)] \geqslant w_x H(x) + \alpha\Pi\,[w_x, H(x)], \quad (14)$$

where H is the total number of employment hours given output, Π is firm profits, w_o is the industry wage, and w_x is the wage demand associated with capacity x (recall $w_x = p/h[x]$). Profits are the difference between revenue and wages, given output:

$$\Pi\,[w_x, H(x)] = px - w_x H(x). \quad (15)$$

Substituting (15) into (14) and rearranging terms yields

$$w_o H(M) \geqslant w_x H(x) - (\alpha/1 - \alpha)p(M - x). \quad (16)$$

The old solution for the optimal productivity schedule had the property that $w_o H(M) = w_x H(x)$ for all x greater than Q^*, less than M. The old solution thus satisfies (16) as an inequality everywhere. This implies that productivity can be raised and total wages lowered without violating the new constraint. Therefore, union ownership of the firm's shares permits greater efficiency in production. This result does not depend on the union's specific objective function.

Union ownership reduces the deterrent necessary to control wage demands. However, the need to maintain some inefficient capacity is

25. See Klein et al. (1978) for a detailed analysis of vertical integration as a response to potential ex post opportunistic behavior. The case of durable special purpose capital analyzed in this paper is precisely the situation which gives rise to quasi rents (in Klein et al.'s terminology); these in turn create the potential for predatory behavior (ex post opportunism) when one party can expropriate the compensation of the other. However, in the case of union-management conflicts, vertical integration is limited by the fact that there is no readily available means for investors to purchase claims to a labor union's resources (see text).

eliminated only in the extreme case of $\alpha = 1$, that is, when the union owns 100% of the firm. Actually, conflicts between the union and other shareholders would be likely to arise for α's substantially below 100%. As the workers' shareholdings approached 50%, the union would come to exercise control over the firm. The union would then be dealing with itself on both sides of the collective bargaining table. At this point, minority shareholders could do little to prevent a settlement unfavorable to their interests.

Finally, consider the case where the restriction preventing investors from buying shares in the union is removed. Union members, in addition to purchasing a fraction of the firm's shares could sell (or pledge) a fraction β of their future wages. Constraint (16) then becomes

$$w_o H(M) \geq w_x H(x) - \frac{\alpha}{1 - \alpha - \beta} p(M - x). \tag{17}$$

Given fixed α and $\beta > 0$, the final term in (17) has greater absolute value than in (16): This means that a function $h(x)$ that causes (16) to be an equality will cause (17) to be satisfied as a strict inequality. A further improvement in productivity is then possible without violating the constraint.

Under the new assumptions, inefficient capacity can be eliminated as long as the workers' share of profits is greater than the fraction of wages they retain: Share of profits $= \alpha \geqslant 1 - \beta =$ retained share of wages. Setting $\alpha = 1 - \beta$ and rearranging terms in (17) it is apparent that this solution makes workers' net compensation (profits plus wages) independent of the actual wages: Whatever wage rate is negotiated, workers receive a fraction of total revenue pM. Clearly, if workers are compensated on the basis of industry revenue instead of on the basis of hours worked then the choice of technology (productivity) will be irrelevant to the bargaining process.

In summary, vertical integration through workers' ownership of firm shares decreases workers' incentives to demand higher wages and permits greater efficiency in the production process. However, the need for inefficient capacity is formally eliminated only in certain limiting cases. In one case the limit corresponds to a worker-owned firm; in the other, workers' net compensation is set at a fixed percentage of total revenue, independent of hours worked.

IV. Conclusion

This paper argues that a market failure exists when capital is durable and specialized and the supply of labor is monopolized by a union. Section I showed that investments justified on the basis of standard capital market criteria will not be undertaken if the capital is durable and the stream of capital returns is vulnerable to expropriation by a union. Section II considered the alternative of investing in technologically diverse capital, that is, in facilities with different labor inputs per unit output. It was shown that in the absence of enforceable contracts investors could do better by combining efficient and inefficient capacity than by investing in efficient capacity alone.

Although serving to protect investors' returns, the so-called productivity strategy is costly to the industry and to society. First, it necessarily raises the average cost of output and thus is feasible only if the firm has market power, at least within a domestic economy. Second, the need to maintain high cost capacity results in reduced incentives to invest in new capacity or in laborsaving technology. Thus, while it may permit some investment that would not take place otherwise, the strategy of deterrence through inefficiency results over time in an industry that is not cost competitive, that grows slowly or not at all, and that has little incentive to innovate or implement new laborsaving technology.

Given these costs, there are clear incentives to workers and investors to reach a cooperative arrangement that makes inefficient plant unnecessary. Section III considered some of the alternatives and related this problem to other problems in contract enforcement. It was shown that the productivity strategy is a special case of contract enforcement through termination of repeated purchase. The purpose of inefficiency in this context is to make the threat of termination (partial shutdown) credible. By raising marginal costs at some facilities management is bound to shut these plants down if a wage demand causes marginal cost to exceed price. Raising marginal cost is costly to investors but may be rational in order to avoid the greater cost of expropriation of capital returns by the union.

Also considered in Section III were third party enforcement (through binding arbitration or debt), extending workers' horizons, and vertical integration. Debt has the advantage of being a unilateral strategy but, it was argued, offers only a partial solution to the problem. In contrast, explicit contracts and binding arbitration can eliminate inefficiency but require that both parties consent in advance to be bound by a contract. In the short run, a predatory party might gain more by refusing to be so bound. The case of the U.S. steel industry was cited as an example of the time and cost involved in waiting for common interests to prevail.

Vertical integration, like debt, offers a partial solution to the problem. Any amount of worker ownership of shares or profit sharing increases the common interests of workers and investors and reduces the amount of inefficient capacity needed as a deterrent. However, within the limits of the formal analysis, the need to maintain *some* inefficient capacity disappears only if workers own 100% of the firm or if their compensation is set at a fixed percentage of total industry revenue.

Finally, in cases where the industry wage is higher than the competitive wage (so that the right to work has positive value), the unions' effective horizon can be extended by allowing current union members to sell their jobs when they quit the industry. In general, there is always some industry wage that, combined with this policy, makes the wage contract self-enforcing; furthermore, in this instance, self-enforcement can be achieved without costly inefficiencies in production.

A comparison of the various contractual alternatives thus reveals that incentives which directly extend workers' horizons by giving them an interest in the future of the industry offer the most promising basis for a stable, efficient, and self-enforcing equilibrium. In contrast, the strategy of deterrence through inefficiency, if it is employed at all, is more likely to represent a transient and pathological industry state. However, even temporary recourse to a defensive strategy involving inefficiency and underinvestment can impose large and unrecoverable opportunity losses on an industry. Understanding this, investors and unions should be encouraged to seek out other means of resolving their contractual differences.

Appendix A
Proof of Proposition 3

Propositions 1 and 2 are implied by (8) and (8a). Furthermore, a function h satisfying (8a) is necessarily monotonically increasing. Therefore the conditions do not violate any known properties of an optimal schedule.

a) By assumption, a productivity schedule satisfying (8) and (8a) makes the wage bill invariant to the wage demanded. In other words, for every wage demand w_x associated with $x \in [Q^*, M]$, labor's gain from higher wages is always just offset by the loss from shutdown capacity.

Assume such a schedule exists. The question is, Can it be improved on? The essence of the proof is to show that deviations from the proposed schedule cannot decrease wages without violating a constraint. To see this, consider figure 7 and assume the schedule *WW* satisfies (8) and (8a). Given x, the loss from wage demand w_x is the area under the curve *WW* to the right of x; the gain

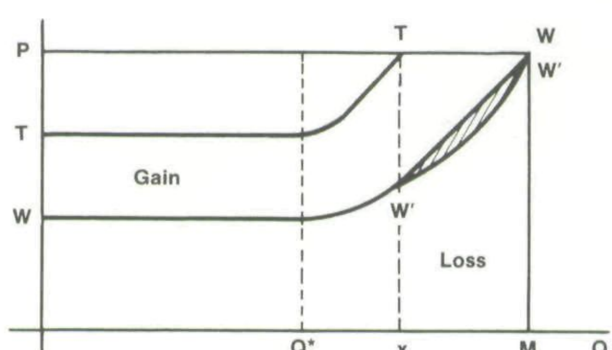

FIG. 7

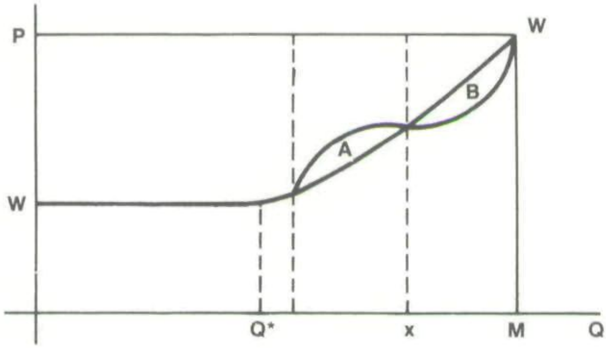

FIG. 8*a*

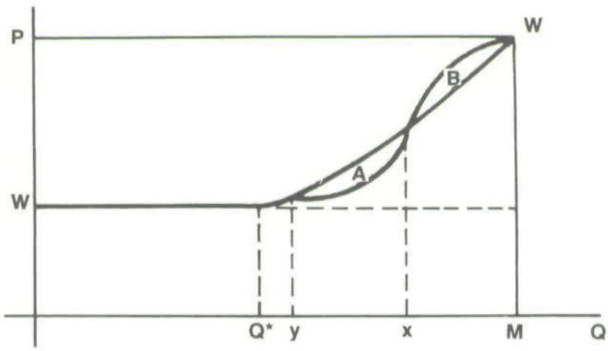

FIG. 8*b*

is the area between the translated schedule $TT = [p/h(x)] \int_0^x h(Q)dQ$ and WW to the left of x.

Consider deviations from the schedule WW. Deviations which monotonically increase h (decrease productivity) on $[Q^*, M]$, increase wages and decrease profits. Deviations that decrease h (increase productivity) violate constraint (8a), because for some x (including $x = Q^*$) the loss is decreased while the gain remains the same. For such a schedule $G_x - L_x \geq 0$ and labor will have an incentive to demand a higher wage (see fig. 7).

Now consider deviations from WW which are positive on some interval and negative on another. Figures 8*a* and 8*b* depict two possibilities. In figure 8*a*, if area A is greater than area B, wages are higher than under WW. If $A = B$, then at the point x, the net gain to the union from wage demand w_x is positive, violating constraint (7b):

$$
\begin{aligned}
G_x - L_x &= (w_x - w_o)\left[h_o Q^* + \int_{Q^*}^{x} h(Q)dQ + A\right] \\
&\quad - w_o\left[\int_x^M h(Q)dQ - B\right] \\
&= w_x A > 0.
\end{aligned}
$$

For $B > A$, $G_x - L_x$ is even higher. Deviations from WW of the type depicted in 8*a* are not feasible.

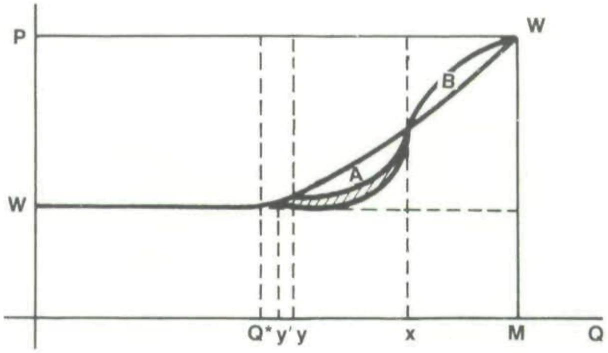

FIG. 9

Now consider figure 8b and assume, for the moment, that area A equals area B. Wages are the same under the old and new schedules, and at point y, $G_y = L_y$ for the new schedule as well as the old. At point x constraint (7b) is satisfied as a strict inequality as can be seen from:

$$G_x - L_x = (w_x - w_o)\left[h_o Q^* + \int_{Q^*}^{y} h(Q)dQ - A\right]$$
$$- w_o\left[\int_{x}^{M} h(Q)dQ + B\right]$$
$$= - w_x A < 0$$

Thus, the new schedule is feasible and as good as the original schedule WW. However, to improve on WW, wages must be decreased; that is, area A must be greater than area B. I assume without loss of generality that A' is as depicted in figure 9. Consider the loss and gain to the union at the point y. The loss under the A' schedule is clearly less than under the A schedule, which is seen immediately from the shaded area in figure 9. It is less obvious but also true that the gain is greater under the A' policy. This follows (after some algebra) from proposition 1 and the fact that by construction the new and old policies diverge on the interval $[y', y]$. A larger gain and smaller loss implies

$$G_y(A') - L_y(A') > G_y(A) - L_y(A) = 0.$$

Therefore the new policy violates constraint (7b). Deviations from WW of the type shown in figures 8a and 8b, although perhaps leading to profits as high as those obtained under WW, cannot strictly improve the policy. More complicated deviations from WW, including discontinuous schedules, can be taken as compound forms of the simple deviations in 8a and 8b.

b) It has been shown that a schedule satisfying (8) and (8a) cannot be improved upon. To prove (b), let $Q^* = M/[1 - \ln(w_o h_o/p)]$ and for $Q \in [Q^*, M]$,

$$h(Q) = \frac{p}{w_o} e^{-(M-Q)/Q^*}.$$

Substituting for Q^* and h it can be verified that (8) and (8a) are satisfied for all x.

c) It was shown above that deviations of the type depicted in figure 8*b* with $A = B$ are feasible and yield the same wages and profits as the candidate schedule *WW*. An infinite number of such alternative optimal schedules can be constructed.

References

Becker, Gary S., and Stigler, George J. 1974. Law enforcement, malfeasance and compensation of enforcers. *Journal of Legal Studies* 3 (January): 1–18.

Council on Wage and Price Stability. 1977. Report to the president on prices and costs in the U.S. steel industry. Washington, D.C. (October).

Federal Power Commission v. *Hope Natural Gas Company*, 320 U.S. 591 (1944).

Heal, Geoffrey M. 1976. Do bad products drive out good? *Quarterly Journal of Economics* 90 (August): 499–502.

Jensen, Michael C., and Meckling, William H. 1976. Agency costs and the theory of the firm. *Journal of Financial Economics* 3 (May): 305–60.

Klein, Benjamin; Crawford, Robert G.; and Alchian, Armen A. 1978. Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics* 21 (October): 297–326.

Klein, Benjamin, and Leffler, Keith B. 1981. The role of market forces in assuring contractual performance. *Journal of Political Economy* 89 (August): 615–41.

Simons, Henry C. 1944. Some reflections on syndicalism. *Journal of Political Economy* 52 (March): 1–25.

Telser, Lester G. 1980. A theory of self-enforcing agreements. *Journal of Business* 53 (January): 27–44.

Telser, Lester G. 1982. The most durable self-enforcing agreement. Mimeographed. Chicago: University of Chicago, Department of Economics. (2d rev., September).