Theoretical Research

Contracts, Wage Differentials and Involuntary Unemployment

Studies in Microeconomics
12(1) 10–31, 2024
© The Author(s) 2024
Article reuse guidelines:
in.sagepub.com/journals-permissions-india
DOI: 10.1177/23210222241245533
journals.sagepub.com/home/mic
Sage

Debraj Ray[1]

Abstract

I study a simple model of labour markets with imperfect monitoring and efficiency wages that provide work incentives. Involuntary unemployment is an outcome, but it is just one particular instance of many inter-firm wage differentials that pervade the equilibrium. The goal of this paper is to study the relationship between firm size, work standards, wages paid, contractual utility for workers, and capital intensity in production. Under certain conditions: (a) larger firms pay higher wages and demand higher individual effort levels; (b) despite these opposing effects, larger firms offer a higher net contractual utility and (c) larger firms are more capital intensive, even when production functions are homothetic in capital and labour.

JEL Classifications: D33, J31, J41

Keywords

Choice under uncertainty, consumer choice, general equilibrium, information economics

Introduction

I dedicate this essay to the memory of my dear friend Ashok Kotwal, whose research has been pivotal for me from the start of my career, and whose gentle and generous demeanour has always inspired me to think about the questions he thought important: the causes and effects of inequality, the nature

[1] New York University, New York, USA

Corresponding author:
Debraj Ray, New York University, New York 10012, USA.
E-mail: debraj.ray@nyu.edu

of informal economic relations, the features of structural transformation, and the interaction of markets and politics. These are some of the questions that I came to be obsessed with, and not by coincidence. More pertinently, this essay is a direct follow-up on Ashok's work.

With Mukesh Eswaran, Ashok wrote several papers (and a book), which I consider to be classical contributions to the field. Among them is the paper, 'A Theory of Two-Tiered Labour Markets,' published in 1985 in the *American Economic Review*. This paper viewed rural labour markets as displaying two separate tiers: *casual labour*, hired seasonally for various observable agrarian tasks (such as harvesting), and *permanent labour*, an occupational niche with long-term arrangements, often used for the supervision and monitoring of casual labour, time-sensitive and hard-to-observe tasks and activities that require exact care. The paper studied different aspects of the institution of permanent labour; specifically, the incidence of such labour as a fraction of the total labour force in agriculture and the root causes of 'casualization' of the labour force. This paper has inspired my own interest in permanent labour from a slightly different angle (Mukherjee & Ray 1995).

My point of departure in this essay stems from a core feature of the Eswaran–Kotwal exercise. This is the idea that casual labour, apart from its own intrinsic significance, can also be viewed as a 'punishment outcome' option for permanent labourers who shirk in their assigned tasks, and are fired. Their result, which permanent workers enjoy a strictly higher utility than their casual counterparts, is a direct implication of this punishment threat. In contrast, in Shapiro and Stiglitz (1984), fired workers enter a pool of unemployed labour, and that serves as a punishment threat just as casual labour does in the Eswaran–Kotwal world. There is a general idea lurking here: that the ability to shirk and the consequent need to monitor a worker and fire him (if need be) can generate an entire hierarchy of occupations with persistent wage differentials between them. In this hierarchy, a fired worker can be absorbed not just into unemployment but into some other job that yields a lower wage. From this perspective, *unemployment* is not necessary for worker discipline, though it might still be necessary in general equilibrium. It is just one 'job category' (the category of not having a job at all), but there can be many other job categories as well. The present paper develops this insight and its implications.

Specifically, I am going to study a simple model of the labour market, with one difference from the textbook competitive framework. The difference is that a labourer will not supply effort unless there are adequate incentives for him to do so. The firm must offer a contract that induces each labourer to put in the specified level of effort.

One approach to this incentive problem is taken in the standard principal–agent model studied by Mirrlees (1975, 1976), Holmstrom (1977), Grossman and Hart (1983) and many others following them. In this approach, a worker's income depends on some *observable* and *contractible* output. By giving the worker a stake in that output, the worker can be made to provide effort that cannot be directly observed.

This is an important contractual form, but it is not the one I consider to be of fundamental interest in the present context. When a potentially large number of workers combine to produce a single output (as in a firm), it is difficult to provide adequate output-based incentives to a sizable fraction of them. For an output-based incentive scheme to have 'power', an additional unit of worker effort must be significantly related to an increase in that worker's income. But with a large number of workers, this cannot be done all around, unless workers can be given large negative payments for some realizations of output, or unless unrealistic 'forcing contracts', which are not robust to the introduction of production uncertainty, are used. For further discussion, see the section on 'Output-Based Incentive Contracts'.

An alternative contractual form, and perhaps more realistic in the context of multi-worker firms, is direct supervision of a worker's effort, coupled with (a) a renewal of the contract should the worker conform to a pre-specified work standard, and (b) eviction or firing should the worker be found shirking relative to this standard. It follows right away that if there is to be any work incentive at all, a worker's utility conditional on obtaining the contract must strictly exceed his utility conditional on being fired.

Suppose, just for the moment, that all firms are identical. Then, in equilibrium, all firms offer the same contractual utility. It follows that the only situation in which this utility can exceed the utility conditional on being fired is one where fired workers cannot find certain reemployment. We conclude that an equilibrium of this model must involve unemployment. Moreover, this unemployment must be involuntary, in the sense that unemployed workers strictly prefer to work, but cannot find a job (Calvo 1979; Eswaran & Kotwal 1985; Shapiro & Stiglitz 1984).

But my main interest is in heterogeneous firms, and in particular the heterogeneity that arises from firm size. As already discussed, my objective is to study not involuntary unemployment per se, but the contractual terms offered by these different firms, and their co-existence and interaction in general equilibrium. Such an equilibrium involves not only the unequal utility treatment of employed and unemployed workers, but also the unequal treatment of identical, *employed* workers across different firms. These differences spring directly from the nature of the 'supervision technology' that I consider. Specifically, I assume that apart from the fixed costs of having supervision at all, 'it is more than twice as costly to supervise two workers than to supervise one.' There is a simple reason for this. A worker's effort is inferred, not always through direct observation of his activity (which is usually difficult and sometimes impossible), but from a number of observed signals that are closely correlated with his true effort. Possibly the most important signal in this regard is the final output produced *by the worker group*. With two workers jointly producing a single output, the information content of the output signal regarding the separate effort level of each individual is significantly reduced. Consequently, to achieve the same level of supervisory accuracy as in the one-worker case, per-worker variable supervision costs must go up.

In formal terms, this assumption is expressed in the postulate that the total variable cost function of supervision is strictly convex in the workforce employed by the firm. For further discussion, see the section on 'Size, Complexity and Hierarchy'. This assumption yields the following empirically testable propositions in the same industry:

1. Larger firms pay higher wages.
2. Larger firms demand higher individual effort levels.
3. Despite the opposing pulls of items 1 and 2 on the worker, larger firms offer a higher net contractual utility.
4. On the other hand, larger firms are more capital intensive,[1] even if production functions are homothetic in capital and labour.[2]

I also consider, informally, a number of extensions of the basic model. These yield further results that are also amenable to empirical investigation. Finally, I briefly consider some possible normative implications of the theory.

All my results stem from the one conceptual premise, discussed above, that makes the model different from the textbook versions. In the present context, Calvo (1979) was possibly the first to conduct a rigorous study of incentive contracts that involve firing. You can also consult Salop (1979), as also the Shapiro–Stiglitz and Eswaran–Kotwal papers I have already discussed. For an integration of the theory of output-based incentive contracts and the contracts discussed here, see Singh (1982) and Dutta et al. (1989).

This essay was first written in 1989 for a Festschrift volume that (for several innocent reasons) never happened. I recall and salute the individual for whom that volume was intended—the distinguished economist Nabendu Sen of Presidency College. Given the obvious indebtedness to Ashok's work, there is excellent reason to believe that Nabendu-babu would approve of the use to which the article is being put today. Some of the material is thoroughly antiquated—I did have to edit references to the 'recent' contributions of Shapiro and Stiglitz (1984) and Eswaran and Kotwal (1985)—but the editors of the current special issue graciously felt that the core ideas still have some value. Therefore, I have let the article essentially stand in its original form, except for several changes in exposition and some variations that do not change the central arguments.

A Labour Market with Firms of Varying Size

I construct the simplest model of a labour market that will allow me to convey the ideas discussed in the Introduction. Suppose that there is a single, homogeneous product produced using capital and labour. Firms hire labour. Their capital endowments are given, and I shall use this as the major variable distinguishing one firm from another.[3] A detailed specification follows.

Labourers

There is a mass of N labourers. Each labourer is taken to be sufficiently small compared to the aggregate, so that I can consider their number as a continuous variable. Each labourer derives utility from income and disutility from work effort.[4] The simplest way of capturing this is to posit that if w is a labourer's income and x is the effort that he puts in, then his net per-period utility is

$$w - c(x),$$

which is aggregated over time using a discount factor $\delta \in (0, 1)$. I assume that

(A.1) c is an increasing, twice differentiable function defined on domain $[0, B)$ (where $0 < B \leq \infty$), with $c(0) = 0$, and $c'(x) > 0$ and $c''(x) > 0$ for $x \geq 0$.[5]

Firms

There is a unit continuum of firms producing a single commodity, the price of which is normalized to 1. The inputs are capital (k) and aggregate labour effort (ℓ). Firms are distinguished only by their capital holdings, which are exogenously given, and in fact, we shall refer to a firm as firm k. Capital ownership in our model is an adequate proxy for firm size. Denote by $F(k, \ell)$ the production function for output. I assume that

(A.2) F has constant returns to scale, it is strictly increasing and smooth whenever $(k, \ell) \gg 0$, it is strictly quasiconcave, and satisfies the Inada conditions in ℓ for every $k > 0$.

We assume that workers must be given appropriate incentives to work. In the textbook principal–agent model, this is achieved to the extent possible by conditioning worker income on the output produced. Here, we study a different incentive structure: the presumption of an effort standard x by each firm, coupled with the supervision of workers and the threat to fire anyone who is found not performing up to standard.

Supervision is costly. This cost is denoted by $\sigma(n)$, where n is the number of workers hired by the firm. In particular, this implies that the cost of supervision does not depend on the chosen work standard. It is unclear whether such dependence is of first order, or even if it is, what its sign should be. A higher work standard may be easier to supervise, as its failure may well be more visible. But it may be also harder to supervise if it involves the performance of a greater number of distinct tasks. I avoid this entire business altogether, except to note that, at the very least, our results will safely extend to all supervision costs that are not 'overly sensitive' to x.[6]

For reasons discussed in the Introduction, I am going to assume that the total variable cost of supervision increases 'more than proportionately' with the workforce of a firm:

(A.3) $\sigma(n)$ is increasing and twice differentiable, with $\sigma(0) = 0$, $\sigma'(n) > 0$, and $\sigma''(n) > 0$ for all $n \geq 0$.

Restrictions on Information Flow

A rich set of theories can be developed by making alternative assumptions about just what happens once a worker is caught shirking and fired. Can his employer ensure, or at least influence, his difficulty of finding a new job? Or is the newly fired worker part of an unemployed pool, indistinguishable from any other job seeker who may have been separated from his job for harmless reasons beyond his control? The particular route I take here is the latter. I assume that the flow of information is limited, and that any job seeker looks the same as any other. Put another way, past employment histories are unobserved by prospective employers.

Optimal Responses to Contracts

With the previous remarks in mind, denote by V the expected lifetime utility available to a currently unemployed worker. V is, of course, a discounted sum of per-period expected utilities (which includes the possibility of remaining unemployed). From a firm's viewpoint, V is an exogenous object. But from an economy-wide perspective, V is an endogenous variable, the determination of which we discuss in the sections entitled 'Lifetime Payoffs' and 'Equilibrium'.

A *contract* is a (time-stationary) pair (w, x), where w denotes income and x is the work standard that the worker is expected to maintain in the job. If the worker fails to meet the standard, he is fired. Otherwise, he is retained on the same terms. The same stationary process continues indefinitely, though I suppose that there is an exogenous probability $q > 0$ that a worker might quit or be removed from his job because of factors not explicitly modelled here. This assumption is merely a shorthand to guarantee some turnover in equilibrium, even when no one is shirking.

The supervision technology is presumed to be accurate, in the sense that shirkers are detected with probability one. This simplifies the analysis because it effectively leaves a worker with two choices: to exactly uphold the work standard or to fully shirk by setting effort level equal to zero. The tweaks needed to accommodate imperfect supervision are discussed in the section on 'Probabilistic Supervision and Uncertain Detection'.

If the worker shirks, then his lifetime utility conditional on being offered the contract (w, x) today is

$$S(w) = w + \delta V, \tag{1}$$

independent of x. The first term on the right-hand side of (1) gives the worker's current payoff w, derived from no effort and the contract income of w. But he is then fired for sure, and from the next period onwards, he receives lifetime utility V, discounted by δ from today's perspective.

The non-shirker, on the other hand, enjoys an expected utility of

$$V(w, x) = [w - c(x)] + \delta(1 - q)\max\{S(w), V(w, x)\} + \delta q V. \tag{2}$$

The right-hand side of (2) has three sets of terms. The first set gives the worker's current utility, taking into account his disutility from maintaining the specified work standard x. As a non-shirker, he retains his job with probability $1 - q$, and then continues to enjoy utility equal to the maximum of shirk/non-shirk utilities. This is discounted by δ, and weighted by its probability of occurrence $1 - q$. With probability q, he loses or gives up his job for exogenous reasons, and then gets V, discounted by δ. The worker will not shirk if and only if $V(w, x) \geq S(w)$,[7] or equivalently if

$$c(x) \leq \delta(1 - q)\left[\frac{w - c(x) + \delta q V}{1 - \delta(1 - q)} - V\right] = \delta(1 - q)\,[V(w, x) - V] \tag{3}$$

The right-hand side of (3) provides the incentive in terms of lifetime utility differences conditional on renewal and expulsion. This must outweigh the disutility of conforming to the specified work standard $c(x)$, which explains the inequality in (3).

Optimal Contracts

Each firm designs its contract keeping an eye on the constraint (3).[8] No more is required: If (3) is met, then the participation constraint $V(w, x) \geq V$ is automatically implied.

In the textbook model of a labour market, firms take the going wages elsewhere as given. Here, firms take the going utility V as given. What does this mean? Of course, our model will not generate a single equilibrium wage for labourers (as you have already guessed from the introduction), and not even a single work standard. The lifetime utility V of a currently unemployed worker will be a mix of his per-period reservation utility, and the utilities from all the contracts available on the market (conditioned on the probability of obtaining these contracts). This utility represents the implicit threat that a firm can impose on a worker by firing him. If the terms and conditions of a firm's offer represents the 'carrot', then V represents the 'stick'.

Now consider a firm's contract design problem. Firm k with capital stock $k > 0$ seeks to maximize its profits by choosing a wage $w(k)$, a work standard $x(k)$ and a workforce $n(k)$. If the standard is adhered to by its workforce, aggregate labour effort is $\ell(k) = n(k)x(k)$. To make sure that work standards are upheld, the firm must incur a supervision cost of $\sigma(n(k))$ and respect the constraint (3) (and then the participation constraint will be automatically met so we can take it that all offers of employment made by the firm will be accepted).

Formally, the firm solves

$$\max_{w,x,n} F(k, nx) - wn - \sigma(n), \tag{4}$$

subject to the constraint (3). We will presently study the characteristics of the solution to this problem in more detail. For now, you can safely presume that for each value of the unemployment utility V, there exists a unique solution to (7). At that solution, we can also suppose that workers will abide by the required work standard $x(k)$ because the incentive constraint (3) has been respected in the contract design.

Lifetime Payoffs

Suppose that firm k offers a stationary contract $(w(k), x(k))$, which is perceived to have expected lifetime utility $V(w(k), x(k))$ and employs $n(k)$ workers. Let $n \equiv \int_k n(k)g(k)dk$, where g is the (exogenous) density of capital ownership. This can only be feasible if $n \leq N$. The remaining workers of measure $N - n \geq 0$ are left unemployed and are 'carried over' to the next date. All this is accounting. Rationality constraints are yet to be imposed, but they soon will be.

With these considerations in mind, we can determine the expected lifetime utility V of a currently unemployed worker. At the start of any date, the pool of unemployed workers is assumed to equal the number of job seekers; there is no on-the-job search. It has two components:

(1) People who have become unemployed by virtue of the exogenous quit rate or because they did not conform to work standards in the firm that employed them. In our model, the latter will not occur because the design of contracts respects (3), and supervision is accurate by assumption. With that rationality constraint imposed, the total number of people in this component is qn.

(2) People who were unable to find employment in the previous period. There are $N - n$ of them.

The total number of vacancies at any date is given by qn. Therefore, the probability that a currently unemployed person will *receive no job offer* is given by

$$\pi = \frac{N-n}{qn+N-n} = \frac{u}{q(1-u)+u}, \tag{5}$$

where $u = (N - n)/N$ is the unemployment rate. Conditional on *some* offer, that offer will be from a firm of type k with probability (density) $n(k)g(k)/n$. Because we know that every firm meets participation constraints, we can take it that all offers will be accepted and all vacancies filled. We must conclude that V, the lifetime expected utility of a currently unemployed worker, is given by

$$V = (1-\pi)\left[\int_k \frac{n(k)}{n} V(w(k), x(k))g(k)dk\right] + \pi[r + \delta V],$$

where $r > 0$ is the per-period reservation payoff of an unemployed worker. Equivalently,

$$V = \frac{1-\pi}{1-\delta\pi}\left[\int_k \frac{n(k)}{n} V(w(k), x(k))g(k)dk + \pi r\right]. \tag{6}$$

It should be reiterated that π, the probability of no offer, is an endogenous variable, and it is positively related to the unemployment rate as in (5).

Equilibrium

An equilibrium is a going utility V^*, a collection of firm contracts $\{w^*(k), x^*(k)\}$ and firm-level employments $\{n^*(k)\}$ for each firm type k, and an aggregate unemployment rate u^*, such that the following three conditions hold:

1. *Aggregate Feasibility*: Defining $n^* \equiv \int_k n^*(k)g(k)dk$, u^* equals $(N - n^*)/N$ and is non-negative.
2. *Profit Maximization*: For every firm k, $\{w^*(k), x^*(k), n^*(k)\}$ solves (7), subject to the incentive constraint (3) in which V is replaced by V^*.
3. *Payoff Consistency*: Defining π^* from u^* as in (5), V^* satisfies (6) with π replaced by π^*.

Our overall model should be consistent. It is.

Proposition 1. *An equilibrium exists.*

Proofs of this and other propositions are in the Appendix. It is unclear whether our assumptions guarantee the uniqueness of equilibrium.

We are now in a position to examine the qualitative features of the model.

The Role Played by Involuntary Unemployment

We begin by observing that the stock of unemployed people, not counting the flows in and out due to labour turnover at the rate qN, is always positive in equilibrium.

Proposition 2. *The equilibrium unemployment rate u^* is strictly positive.*

There is a simple intuition for this result. If there were full employment in equilibrium, then the going utility would be expressible as a convex combination of the various utilities available in the market, with 'unemployment utility' r getting zero weight. Notice how *any* firm k that asks for a strictly positive work standard must actually offer worker utility $V(w^*(k), x^*(k)) > V^*$ – the participation constraint cannot bind if the incentive constraint (3) is to hold. In our model this is indeed true of every firm. But then we have a contradiction: V^* cannot be a convex combination of the $V(w^*(k), x^*(k))$ values alone.

As we have just noted, our model assumes that *every* firm type must incentivize work effort. In reality, one can imagine a number of activities, such as self-employment, or production using labour with outputs clearly and contractually attributable to each individual, or production in which there is on-the-job supervision with wages paid ex post, for which effort provision does not require dynamic incentives. For each such activity, the participation constraint binds and there is no separate incentive constraint. If that were true of every activity, we would be back to an introductory textbook model of the labour market. However, as long as there are *some* jobs that require incentivized effort using the dynamic threat of termination, the argument above holds as long as labour turnover from such firms is positive. Unemployment must appear in equilibrium.

Now that sounds paradoxical. You might respond: Unemployment is just an activity, formally like any other. Why would we specifically need *this* activity to co-exist with the others? The answer is that unemployment can be viewed as an activity, yes, but it is special in that it allows for job search while engaged in that activity. If we were to extend our model even further, so that individuals can search for jobs while engaged in self-employment or in activities without dynamic effort constraints, then the model *does* become compatible with full employment. On the other hand, the feature of unequal treatment of identical workers across jobs persists. I discuss this last observation in more detail below.

Another substitute for the necessity of unemployment comes from worker 'scarring'. If a fired worker is known to have been dismissed from a job, then his future lifetime payoff could be made dependent on the job he had to begin with. For instance, in the Eswaran–Kotwal model, there are just two occupations – casual labour and permanent labour. The former activity requires no effort incentives, by assumption. But if there is shirking in the latter, a fired worker is absorbed into the casual labour force for good, and there is no returning to the permanent labour force. This is an instance of worker scarring, and it obviates the need for open unemployment. (Unequal treatment across jobs continues to be necessary.) The necessity for unemployment also hinges, then, on weak information flows: A worker in the unemployed pool has no history attached to him.

An unemployed person is *involuntarily unemployed* if he strictly prefers to work in one or more of the available jobs, and if these jobs employ individuals who are in all respects identical to him. Involuntary unemployment, in its broadest sense, deals with unequal labour-market treatment of ex-ante identical individuals.[9] A particularly strong form of involuntary unemployment is true in our setting.

Proposition 3. *In any equilibrium, all unemployment is involuntary. Indeed, in this model, we have the stronger result that* $V(w^*(k), x^*(k)) > V^*$ *for every* $k > 0$.

It would have been enough for involuntary unemployment to say that there are *some* firms k with $V(w^*(k), x^*(k)) > V^*$. After all, all workers are identical and this would suffice for unequal treatment. But in our model, a stronger observation applies: The previous inequality holds for every firm. The dynamic incentive constraint implies that the holding of a job confers strictly higher utility than the state of being unemployed. Be careful, though: from this observation, no deduction should be drawn regarding the *social* desirability of unemployment as a means of creating work incentives. It is a positive description of the equilibrium outcome of a profit-maximizing, competitive economy. And, I might add, it is a description that is sharply at variance with the essentially harmonious view of competition as envisaged by Adam Smith and his followers. See, however, the section 'Some Normative Remarks' for some remarks on the possible normative implications of the theory.

Shapiro and Stiglitz (1984) have already highlighted this aspect of unemployment. But the core aspect I wish to emphasize here is more closely connected to Eswaran and Kotwal (1985). Drawing on their idea of casual labour as a punishment device, suppose in fact that the reservation utility of our model is generated, not by unemployment, but by some productive activity which is not costly to supervise, such as self-employment. With this convenient re-naming, and allowing for on-the-job search under self-employment, it is easy to see that our model will still generate employment for some workers whose contractual utility is strictly higher than that of the self-employed, and unemployment is unnecessary. In this sense, it is not unemployment that is a core feature, but rather the fact there are utility differentials among identical individuals – employed or not – that persist in equilibrium. Among these, the state we call 'unemployment' involves no incentive constraints of its own, *and* allows for search. In the rest of the paper, I focus on these inter-firm differences, and push the question of involuntary unemployment into the background.

Contract Differentials and Firm Size

I have already observed that, in general, there will be wage and utility differentials in equilibrium among workers who are identical to start with. The purpose of this section is to demonstrate that such differentials are related in a systematic way to the size of the firm, and to outline a number of extensions to the basic model in this context.

Differentials

By a larger firm, I shall mean a firm with a larger endowment of capital. A larger firm naturally wishes to engage a larger workforce, given that labour is complementary to capital. At this point, the technology of supervision becomes a critical factor, Recall our assumption that the cost of supervision is convex in the labour force. This convexity forces the larger firm to take steps other than a straightforward expansion of its workforce. Specifically, we have:

Proposition 4. *In equilibrium, a larger firm (i) sets higher work standards ($x^*(k)$ is increasing in k) (ii) offers a higher wage ($w^*(k)$ is increasing in k), (iii) offers a contract with a higher worker payoff ($V(w^*(k), x^*(k))$ is increasing in k) (iv) hires a larger workforce ($n^*(k)$ is increasing in k), and (v) is more capital-intensive in the sense that $\frac{k}{n^*(k)x^*(k)}$ — and a fortiori $\frac{k}{n^*(k)}$ — are increasing in k.*

In particular, a worker will strictly prefer to work in the firms with highest capitalization, which are also the most capital-intensive.

Here is the main idea behind the proposition. Supervision is costly, and proportionately costlier as a firm expands. There is, therefore, a tendency to not expand employment in proportion to capital ownership, even though employment does expand in absolute terms (part iv). To compensate for this in some measure, the work standard for each labourer is raised (part i). Despite these two factors which both act to increase total labour effort, the net effect is always an increase in the capital-to-total-effort ratio, and certainly an increase in the capital-to-employment ratio (part v).

The rest of the details follow without much fuss. Once work standards are raised, a larger differential between contractual utility and the unemployment utility V^* is needed to satisfy the worker's incentive constraint. This explains part iii. Of course, as a result of this, the wage offered must be higher, which implies part ii.

These observations might explain, at least partly, why larger and more complex organizations tend to pay higher wages for similar jobs, and why individuals prefer these jobs. Such organizations also appear to demand higher work standards. Of course, the presence of these differentials is an empirical question. I hope the theory is provocative enough to induce such empirical analysis.

It must be added that this theory seeks to explain differentials across small and large firms, not by taking recourse to assertions that larger firms and/or their workers are intrinsically more 'efficient.' That may well be the case, and it could account for why those firms are large to begin with. It is a perfectly reasonable alternative explanation. However, what I am arguing is that there may be a more fundamental reason why the 'law of one price' may not hold in the case of labour. The reason stems from the necessity to supervise labour effort, and from the fact that supervision is costly. Larger firms react to this by cutting employment (in relative terms), and by demanding greater effort from each employee. The price paid for this demand is a higher wage; indeed, a higher worker utility. After all, this is what gives the threat of firing its credibility.

Having unabashedly promoted Proposition 4, I must now qualify it. It is possible to overturn this result in some scenarios. For instance, suppose that supervision costs also depend on work standards, and there is a strong and positive complementarity between those work standards and the marginal supervision cost as employment changes. Then a larger firm, faced with the prospect of a larger labour force, might face an extremely steep supervision bill as it seeks to expand employment. It might be better off bringing down work standards in the process, leading to a reversal of our results. Do I find this scenario compelling? The answer is no. As mentioned before, I am unconvinced that supervision costs should depend on work standards to begin with, let alone that they should positively affect the marginal supervision cost of additional employment. At the same time, I do want to convey the point that Proposition 4 does implicitly or explicitly rest on a set of assumptions that you, the reader, might want to explore more critically.

Discussion and Possible Extensions

Size, Complexity and Hierarchy

The nature of the supervision technology is related to the 'complexity' of the production activity; specifically, the degree of interpersonal interaction required to produce the final team product. I have proxied this complexity by firm size, and I do not believe it is a bad proxy. Firms employing more capital and labour in the same industry are likely to be more complex: if there is any deficiency that we can point to in aggregate performance, it is that we are closer to the boundaries of the span of control. It is more difficult to hold particular individuals responsible for a failure. Flipping this argument, we can conclude that the precise monitoring of an individual's performance becomes a more costly activity in larger firms.

This discussion, and the model in general, throws some light on the question of *intra*-firm organization. These issues have been explored by Calvo and Wellisz (1979), but there is room for further research. Hierarchy implies 'responsibility in layers.' After a certain critical level of the workforce is reached, it may be profitable for the firm to hire a divisional manager who takes responsibility for a clearly identified subdivision of the workforce, This manager is given a contract, just as the workers are, and is fired if something goes wrong in the activity of that subdivision. As in Calvo and Wellisz (1979), one can construct realistic models to show that managers will enjoy a higher net utility than the workers he takes charge of, even though there may be no intrinsic differences in manager-worker ability. With a larger firm, it may be profitable to hire a manager to take charge of the divisional managers, and so on, leading to longer chains of wage differentials.

If one takes into account these additional features and reconstructs the general equilibrium model of the preceding sections, there will be a significant enrichment of (and – no doubt! – some differences in) the results. This is as it should be. The power of our approach lies not in the detail of individual results, but in its basic conceptual postulate and some of its striking implications.

Probabilistic Supervision and Uncertain Detection

In my model I have assumed – quite unrealistically – that supervision is carried out in every period, and that shirkers are detected with probability one. Both assumptions can be dropped to achieve extensions of some interest. Consider, first, the new decision problem of the firm in a model where the first assumption is dropped. The firm chooses (w, x, n) as before, and now additionally p, some stationary probability of supervision, to solve

$$\max_{w,x,n,p} F(k, nx) - wn - p\sigma(n, x), \tag{7}$$

subject to a constraint analogous to (3) — the supervision probability must now be included in the constraint. I am assuming here that the firm can credibly commit to such a probability.[10]

We can go ahead and define an equilibrium just as we did before. What would we find? I do not know for sure (the results need to be worked out), but I would conjecture that exactly for the same reasons advanced at the beginning of this section, the probability of supervision would be lower for the larger firms. Such firms must exploit every instrument they can to escape high supervision costs, and probabilistic supervision is precisely one such instrument. Of course, if this conjecture is borne out, it would further reinforce the utility differentials observed in Proposition 4.

Now let us drop the second assumption, which states that shirkers are detected with probability one whenever they are supervised, One way to do this is to introduce a 'probability-of-detection' for the worker, $p(x, x')$, which depends on x, the stated work standard, and on x', the actual work effort put in. Such a function will modify the incentive constraint (3). With this modification, the main difference is that in equilibrium, some shirking will occur and some proportion of workers will be fired. In my model, this feature is absent except via the exogenous quit component q.

Introducing a probability of detection opens the door to another source of higher wages in larger firms. The complexity of a larger organization and the greater intertwining of joint production implies that in such organizations, the probability of detection is lower. That tightens the incentive constraint, because the worker has greater incentives to shirk if he has a lower probability of being caught doing so. To restore balance, it will generally be necessary to increase the efficiency wage.

Output-Based Incentive Contracts

In the formal analysis, I have considered only one type of labour contract. It provides incentives by driving a wedge between the contract utility and the 'going utility,' and follows up by threatening to fire shirkers. I have already noted that the 'classical' contract studied in the standard principal-agent literature is different. Specifically, the income payments to a worker are related to the output produced.[11] There is no reason why a contract in our model should be completely devoid of output-based incentives.

However, in the particular context of the model being studied here, one observation is quite clear. With a single, jointly produced output, output-based contracts lose their power as the workforce expands. After all, it is not possible to simultaneously provide a large number of workers with a significant share in the output. I am therefore led to the following conjecture: in tasks that are jointly performed by a large number of workers, one would expect to observe contracts that combine explicit supervision with the threat to fire workers. Output-based incentives would occupy a secondary position here. Conversely, output-based payment schemes would acquire a greater role in tasks involving a relatively small number of workers, though firing clauses would not be absent, in general.[12] The contrasting implications for the same productive activity in large versus small firms should now be quite clear.

Some Normative Remarks

Ours is a positive theory, in that it purports to explain some aspects of observed reality, and makes ancillary predictions that may require further empirical analysis. Are there any normative lessons to be learnt?

There are, in fact, some normative issues of unemployment policy that emerge. These are addressed in Salop (1979) and Shapiro and Stiglitz (1984), and I will avoid repeating these points.

Rather, I remark briefly on some implications for the design of contracts. I shall motivate these by taking up an old issue: the question of efficiency differences between private and public sector firms. I do not, by the way, treat such differences as an empirically established fact, but only as a feature that appears to be valid on the basis of casual observation. These differences coexist with the observation that private firms appear to pay more than their public sector counterparts for labour that is similar. Our model has, of course, an obvious explanation for this phenomenon, which is that greater flexibility of the private sector in firing decisions permits a higher work standard. At the same time, in order to satisfy the resulting incentive constraint, private sector firms must pay more.

The inability to fire shirkers in the public sector arises partly from a normative consideration: that stability of employment 'should' be guaranteed in a labour-surplus economy.[13] Is there any ethical justification for this? This is a difficult question, given that the very same stability restricts job opportunities for the unemployed. However, let us grant for the sake of argument that, at least for low-income, unskilled jobs, employment stability should be a consideration. In that case, we must be prepared to bear its inevitable consequence: that such workers will exhibit dramatically lower productivity relative to their private sector counterparts.

What about higher-level employees? The ethical considerations of job stability must surely be weaker, if not absent. Yet a comparable degree of insulation persists at these levels as well. We cannot simultaneously bemoan the fact that there is low productivity, *and* uphold unconditional employment stability. To gain the one is to sacrifice the other, unless we are prepared to believe that public sector employees are somehow imbued with the highest degree of social consciousness.

Summary

Inspired by the work of Shapiro and Stiglitz (1984) on involuntary unemployment and Eswaran and Kotwal (1985) on permanent labour contracts, this paper outlines a simple model of wage differentials (across small and large firms). The analysis is based on the premise that each worker in a firm must be supplied with an appropriate reward/punishment incentive contract in order to put in effort. This contractual structure is achieved by offering the worker a lifetime payoff that *strictly* exceeds his lifetime utility conditional on unemployment. The offer is backed up by rewarding compliers with contract renewal, and shirkers with expulsion.

With no on-the-job search, an economy based on such contracts must exhibit involuntary unemployment in equilibrium, and will generally display wage differentials across firms of different sizes. At the same time, the very feature that a multitude of jobs exist, with their attendant wage differentials, highlights the observation that unemployment can formally be viewed as 'just another job.' The payoff difference between unemployment and other jobs can then be interpreted as *involuntary* unemployment. But it also brings out the special role played by the 'unemployment job': that it allows for 'on-the job' search. If such searches were available on other jobs with no supervision problems, open involuntary unemployment would not be necessitated in equilibrium.

But my purpose is to go beyond these formal parallels and tease out some testable implications of the model. I do that by examining the nature of wage differentials across firms of different sizes. I show that firms with larger capital ownership will pay a higher wage, and set a more demanding work standard. The combination of these factors appears ambiguous at first glance, but in fact, such contracts must provide a higher net payoff to a worker. The model additionally predicts that larger firms will tend to be more capital-intensive, even if the production function is homothetic in capital and labour. These outcomes are consequences of the technology of supervision, and of the way in which this technology is affected by the size and complexity of the organization. A number of extensions and some normative implications are explored.

This paper was written over three decades ago, but I believe that its approach remains relatively unexplored. A recent literature on relational contracts has come into full maturity, with many authors exploring the rich variety of within-firm relationships that can result from such considerations. ‘Relational contracts’ refer to the idea that some objects, such as effort, might be *observable* but not *contractible*. The distinction is of great importance. The absence of contractibility means that we cannot write formal contracts that link current compensation to some observables. But that does not stop us from conditioning on those observables when deciding whether or not to *renew* a contract. My paper falls into that category of models. But its particular emphasis is yet to be fully incorporated into theories of the labour market. I refer to the ‘general equilibrium’ of relational contracting: the maze of macroeconomic price and quantity adjustments that allow an economy-wide system of relational arrangements to lock together. Along with Shapiro–Stiglitz and Eswaran–Kotwal, this paper represents a very preliminary exercise in that broad area.

Appendix

Consider the problem set out in the text:

$$\max_{w,x,n} F(k, nx) - wn - \sigma(n),$$

subject to the constraint (3), which can be equivalently rewritten as

$$c(x) \leq \delta(1-q)\,[w-(1-\delta)V]\,. \tag{8}$$

Observe that for any choice of x, w must be chosen so that (8) holds with equality; that is,

$$w = \frac{c(x)}{\delta(1-q)} + (1-\delta)V.$$

Substituting this into the maximand, we generate the following maximization problem in (x, n):

$$\max_{x,n} F(k, nx) - \left[\frac{c(x)}{\delta(1-q)} + (1-\delta)V\right] n - \sigma(n). \tag{9}$$

By our assumptions, a maximum for (9) must exist. Moreover, by the Inada conditions on the production technology, it must be that a maximum involves $(x, n) \gg 0$, and so must satisfy the interior first order conditions

$$\delta(1-q)F_\ell(k, nx) - c'(x) = 0 \tag{10}$$

with respect to x and

$$\delta(1-q)\left[xF_\ell(k,nx)-\sigma'(n)-(1-\delta)V\right]-c(x)=0. \tag{11}$$

with respect to n.

To verify that only a single pair (x, n) can satisfy these equalities, it is sufficient (using a familiar index number argument; see, e.g., Varian 1975) to check that the determinant of the relevant Jacobian at any solution to (10) and (11) is positive. Suppressing arguments for ease in writing, the Jacobian is easily seen to be

$$\mathcal{J}=\begin{bmatrix} \delta(1-q)nF_{\ell\ell}-c'' & \delta(1-q)xF_{\ell\ell} \\ \delta(1-q)\left(F_\ell+nxF_{\ell\ell}\right)-c'(x) & \delta(1-q)\left(x^2F_{\ell\ell}-\sigma''\right) \end{bmatrix}$$

The determinant of this object *at any zero* of the system (10) and (11) has the same sign as

$$D\equiv\det\begin{bmatrix} \delta(1-q)nF_{\ell\ell}-c'' & \delta(1-q)xF_{\ell\ell} \\ nxF_{\ell\ell} & \left(x^2F_{\ell\ell}-\sigma''\right) \end{bmatrix} \tag{12}$$

which is easily verified by direct computation to be strictly positive. It follows that there is a unique solution to the maximization problem (9).

Proof of Propositions 1 and 2. Denote the solution by $\{w(k,V),x(k,V),n(k,V)\}$ for every $k>0$. Define $n(V)\equiv\int_k n(k,V)g(k)dk$, and then

$$\pi(V)\equiv\max\left\{\frac{N-n(V)}{qn(V)+N-n(V)},0\right\} \tag{13}$$

for all $V\geq 0$. With this in hand, define for every $V\geq 0$,

$$\Psi(V)\equiv[1-\pi(V)]\left[\int_k\frac{n(k,V)}{n(V)}V(w(k,V),x(k,V))g(k)dk\right]+\pi(V)[r+\delta V], \tag{14}$$

where we recall that r is the per-period payoff conditional on unemployment. Because firm responses are unique for every k and v, and all responses are upper hemicontinuous by standard arguments, it follows that $\{w(k,V),x(k,V),n(k,V)\}$ and therefore $\pi(V)$ are all continuous in V. By standard arguments using the continuity of the integral in expression (14) in V, we can therefore see that $\Psi(V)$ is a continuous function.

Additionally, because $\pi(0)>0$ and $r>0$, we have $\Psi(0)>0$.

Finally, let us examine the opposite limit as $V\to\infty$. It is easy to see that for every firm, $w(k,V)\to\infty$ and so $n(k,V)\to 0$. Therefore $n(V)\equiv\int_k n(k,V)g(k)dk\to 0$ as well, and so, using (13), $\pi(V)\to 1$ as $V\to\infty$. It follows that for large V,

$$\Psi(V) \approx r + \delta V \ll V.$$

These end-point arguments plus continuity guarantee that there is a fixed point V^* of Ψ; that is, there is $V^* > 0$ such that $\Psi(V^*) = V^*$.

We complete the proof by observing that any fixed point V^* of Ψ generates an equilibrium, where V^* becomes the lifetime utility conditional on being unemployed today, and where we set $w^*(k) = w(k, V^*)$, $x^*(k) = x(k, V^*)$, $n^*(k) = n(k, V^*)$, and define $\pi(V^*)$ from (13) and finally u^* from $\pi(V^*)$ using (5). The definition of an equilibrium in the section entitled "Equilibrium" has three components, and of these, two of them — profit maximization and payoff consistency — are immediately satisfied by construction of the mapping Ψ. It only remains to show aggregate feasibility: that $n(V^*) \leq N$, a condition that is not automatically guarantees from the definition of Ψ.

In fact we shall show the stronger result that $n(V^*) < N$, and therefore that $u^* > 0$, which will also prove Proposition 2. Suppose not; then $n(V^*) \geq N$, and therefore $\pi^* \equiv \pi(V^*) = 0$ from (13). Using this and the fixed point property, we must conclude that

$$V^* = \int_k \frac{n^*(k)}{n(V^*)} V(w^*(k), x^*(k))g(k)dk,$$

but we also know from the incentive constraint (3) that $V(w^*(k), x^*(k)) \geq V^*$ for every k, and so it must be that

$$V(w^*(k), x^*(k)) = V^* \text{ for all } k. \tag{15}$$

Using (3) again along with (15), we must conclude that $x^*(k) = 0$ for all k, a contradiction to the fact that $x^*(k) > 0$ for all k. So we have shown that $n(V^*) < N$, and therefore that $u^* > 0$, which is Proposition 2. Proposition 1 then follows because all the conditions of equilibrium are now met at $V = V^*$.

Proof of Proposition 3. For any k, the incentive constraint (3) implies that

$$V(w^*(k), x^*(k)) - V^* \geq \frac{1}{\delta(1-q)} c(x^*(k)) > 0,$$

because we already know from the Inada conditions for F that $x^*(k) > 0$ for every k.

Proof of Proposition 4. The various results in Proposition 4 are obtained by totally differentiating (10) and (11). For our purposes, it will suffice to consider parametric changes in k and V only. We obtain:

$$[\delta n(1-q)F_{\ell\ell} - c'']dx + \delta x(1-q)F_{\ell\ell}dn = -\delta(1-q)F_{\ell k}dk \tag{16}$$

and

$$nxF_{\ell\ell}dx + \left[x^2F_{\ell\ell} - \sigma''\right]dn = -xF_{\ell k}dk + (1-\delta)dV, \qquad (17)$$

where the second equation has been simplified by using (10) and then dividing through by $\delta(1-q)$. Using Cramer's Rule, we see from (16) and (17) that

$$\frac{dx}{dk} = \frac{1}{D}\begin{bmatrix} -\delta(1-q)F_{\ell k} & \delta(1-q)xF_{\ell\ell} \\ -xF_{\ell k} & x^2F_{\ell\ell} - \sigma' \end{bmatrix} = \frac{\delta(1-q)F_{\ell k}c''}{D} > 0, \qquad (18)$$

where D is given from (12). (To verify the claimed sign, recall that $D > 0$ and that the complementarity $F_{\ell k} > 0$ is a consequence of our assumptions on F.)

This verifies part (i) of Proposition 4. Now recall (10), which tells us that F_ℓ is positively related to x. Therefore, by the linear homogeneity of F, so is k/nx. Then append (18) to this last observation to verify part (v) of the proposition. Next, observe that

$$\frac{dn}{dk} = \frac{1}{D}\begin{bmatrix} \delta(1-q)nF_{\ell\ell} - c'' & -\delta(1-q)F_{\ell k} \\ nxF_{\ell\ell} & -xF_{\ell k} \end{bmatrix} = \frac{xF_{\ell k}c''}{D} > 0. \qquad (19)$$

Expression (19) verifies part (iv) of the proposition.

Return to (18) and apply it to the constraint (3). We immediately see that contractual utility ($w(k)$, $x(k)$) is an increasing function of k. This verifies (iii) of the proposition. And finally, apply (18) to the rewriting (8) of the incentive constraint to see straight away that $dw(k)/dk > 0$. This verifies pari (ii) of the proposition, and so completes the proof of Proposition 4.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author gratefully acknowledges funding from NSF grant SES 2315720.

Notes

1. By capital intensity, we mean the ratio of capital to total effort, that is, the individual work standard multiplied by total employment. In light of item 2, this means that if capital intensity is defined by the capital-workforce ratio, then the capital deepening effect is even more pronounced.
2. In fact, we assume that the production function displays constant returns to scale, capital and total worker effort, so that homotheticity is automatically implied.

3. All the results go through even in the presence of a capital market. But then it will be necessary to introduce alternative sources of heterogeneity, such as differences in productivity.
4. The assumption of work disutility at all levels of work effort is an exaggeration and not meant to cast aspersions on the work ethics of people! The idea is simply that labour must be provided incentives to supply effort over and above some minimum, which is normalized here to zero.
5. The assumption on c'' reflects the presumption that the marginal disutility of effort increases with effort. By the way, if you like to be general and wish to describe the worker's utility function as $u(w, x)$ (with the proper assumptions on derivatives, of course), you can redo the analysis this way with no loss in the results.
6. I am also not going to explicitly consider any fixed costs of supervision. That fixed cost is only needed to examine whether firms will produce or shut down, and this minor extension can be easily accommodated.
7. We presume that if a worker is indifferent between shirking and not shirking, then he will not shirk.
8. It might be argued that firms do not know the form of the disutility function $c(x)$, so they cannot be sure of observing (3) even if they want to. This is a valid objection, but it is one that takes the formal model too literally. The model is an approximation to the more realistic scenario where firms have an imprecise notion of the form of the incentive constraint. After all, to argue that firms have *no* knowledge of it is even more unrealistic.
9. This notion could be problematic, given that no two people are ever exactly identical. But it is easy to extend the concept: unequal treatment can be taken to refer to a discontinuity in equilibrium payoffs as a function of individual characteristics (as in Dasgupta & Ray, 1986). For our model, however, the simpler definition will suffice.
10. Whether such pre-commitment can indeed be credible is an important issue. If not, we must look at equilibria that involve mixing by both principal and agent.
11. Obviously, in agriculture, this is a commonly observed contractual form. Two examples are sharecropping and piece-rate contracts in harvesting.
12. On the coexistence of both types of clauses, see Dutta et al. (1989).
13. There may be other considerations, such as the possible abuse of the power to fire by individuals who possess such power.

References

Calvo, G. (1979). Quasi-Walrasian theories of unemployment. *American Economic Review*, *69*, 102–106.

Calvo, G., & Wellisz, S. (1979). Hierarchy, ability and income distribution. *Journal of Political Economy*, *87*, 991–1010.

Dasgupta, P., and Ray, D. (1986). Inequality as a determinant of malnutrition and unemployment: Theory. *Economic Journal*, *96*, 1011–1034.

Dutta, B., Ray, D., & Sengupta, K. (1989). Contracts with eviction in infinitely repeated principal agent relationships. In P. Bardhan (Ed.), *The economic theory of agrarian institutions* (pp. 93–121). Oxford University Press.

Eswaran, M., & Kotwal, A. (1985). A theory of two-tier labor markets in agrarian economies. *American Economic Review*, *75*, 162–177.

Grossman, S., & Hart, O. (1983). An analysis of the principal agent problem. *Econometrica*, *51*, 7–45.

Holmstrom, B. (1977). Moral hazard and observability. *Bell Journal of Economics*, *10*, 74–91.

Mirrlees, J. (1975). *The theory of moral hazard and unobservable behavior* (mimeo). Nuffield College.

Mirrlees, J. (1976). The optimal structure of authority and incentives within an organization. *Bell Journal of Economics*, *7*, 105–131.

Mukherjee, A., & Ray, D. (1995). Labor tying. *Journal of Development Economics*, *47*, 207–239.

Salop, S. (1979). A model of the natural rate of unemployment. *American Economic Review*, *69*, 117–125.

Shapiro, C., & Stiglitz, J. (1984). Equilibrium unemployment as a worker discipline device. *American Economic Review*, *74*, 433–444.

Singh, N. (1982). *The possibility of nonrenewal of a contract as an incentive device in multiperiod principal agent models* (mimeo). Department of Economics, University of California Santa Cruz.

Varian, H. (1975). A third remark on the number of equilibria of an economy. *Econometrica*, *43*, 985–986.