THE ECONOMIC JOURNAL

DECEMBER 1982

The Economic Journal, **92** (*December* 1982), 787–804
Printed in Great Britain

INCREASING RETURNS AND THE FOUNDATIONS OF UNEMPLOYMENT THEORY*

Martin L. Weitzman

Current approaches to macroeconomic theory divide, roughly, into two basic schools of thought. One approach, typically associated with 'rational expectations', views the economy as basically at or near Walrasian market clearing equilibrium, in which case it becomes very difficult to explain away convincingly the undeniable persistence, at times, of involuntary unemployment. Another approach views the underemployed economy as being in a state of temporary fixed price disequilibrium, in which case, aside from having to repress suspicions that the essence of effective demand failure might not have to do with prices fixed at 'wrong' values, there is a lot of rationalising to be done about why profitable adjustments are so slow. This paper argues that something like an internally consistent theory of steady state involuntary unemployment is possible, and even plausible. The cornerstone of this third approach is increasing returns to scale, which blocks unemployed agents from breaking out of a low level equilibrium trap unless there is overall coordination or stimulation. A simple model is presented which contains the basic ideas.

I. INTRODUCTION: UNEMPLOYMENT AS A FAILURE OF COORDINATION

I should begin by describing the phenomenon I am trying to explain. 'Unemployment equilibrium' is persistent involuntary underutilisation of the major factors of production, caused by insufficient aggregate demand. The market system suffers from a 'failure to coordinate' the desired consumption and production plans of all agents because the unemployed lack the means to communicate or make effective their potential demands.

In a modern economy, many different goods are produced and consumed. Each firm is a specialist in production, while its workers are generalists in consumption. Workers receive a wage from the firm they work for, but they spend it almost entirely on the products of other firms. To obtain a wage, the unemployed worker must first succeed in being hired. However, when demand is depressed because of unemployment, each firm sees no indication it can profitably market the increased output of an extra worker. The inability of the unemployed to communicate effective demand results in a vicious circle

* Without attributing to them opinions or errors in the paper, I would like to express gratitude to R. M. Solow, D. D. Hester, J. Tobin, F. H. Hahn, M. Bruno, A. K. Dixit, and a referee and editor of this JOURNAL for useful comments.

of self-sustaining involuntary unemployment. There is an atmosphere of frustration because the problem is beyond the power of any single firm to correct, yet would go away if only all firms would simultaneously expand output. It is difficult to describe this kind of 'prisoner's dilemma' unemployment rigorously, much less explain it, in an artificially aggregated economy that produces essentially one good.

Unemployment equilibrium as a story about effective demand failure coexists uneasily with classical general equilibrium theory. Most economists deal with both concepts, but in a kind of schizophrenic manner. Unemployment and classical versions of equilibrium theory are not just different approaches; in a fundamental sense they seem to represent almost incompatible paradigms.

Classical general equilibrium theory starts at a very basic level by specifying tastes, technology and endowments. The world is peopled by atomistic economic agents who optimise. They interact impersonally through purely competitive markets. There are no genuine economic organisations. Prices adjust freely and Say's Law guarantees full employment of factors. An elegant and complete answer is given to the 'what, how, and for whom' question. Major issues concern efficiency and the determination of relative prices.

By comparison, effective demand theory appears to lack a thoroughly consistent microeconomic specification. At least implicitly, the effective demand approach involves a behavioural splitting of decision roles among different agents, and the existence of genuine organisations. The key idea is that insufficient aggregate demand will cause output to fall below the full employment level. Without remedial action, the economy can get stuck in underemployment states. Prices play a negligible adjustment role in clearing markets and Say's Law does not operate. Primary emphasis is on policy measures the government can take to correct unemployment.

In their methodology, assumptions, and conclusions, the two approaches represent very different world views. If not mutually exclusive, they are at least difficult to reconcile. All this notwithstanding, in some sense the average mainstream economist pledges allegience to both paradigms. Predictably, that kind of intellectual schizophrenia frequently results in confusion.

If unemployment equilibrium is viewed as a form of steady state general equilibrium theory, why does it differ so radically from the classical version? On the most basic level of tastes and technology, which hidden assumption marks the first dividing line?

In this paper I want to argue that the ultimate source of unemployment equilibrium is increasing returns. When compared at the same level of aggregation, the fundamental differences between classical and unemployment versions of general equilibrium theory trace back to the issue of returns to scale. A crucial distinction is whether or not the underlying technology is convex. Classical equilibrium is essentially organised around the postulate of *constant* returns to scale. The assumption of *increasing* returns to scale provides a basic organising principle for unemployment equilibrium theory.

More formally, I hope to show that the very logic of strict constant returns to scale (plus symmetric information) must imply full employment, whereas

unemployment can occur quite naturally with increasing returns. In this sense, unemployment equilibrium is tied to an underlying technology of increasing returns, just as classical general equilibrium is bound up with constant returns. If that is a fair way of looking at the matter, small wonder inconsistency or confusion can result from trying to apply the concepts and tools of one environment to the other.

The paper presents a simple general equilibrium model – perhaps it is more accurate to call it an example – which clearly indicates the connection between microeconomic increasing returns to scale and macroeconomic involuntary unemployment. A closed, circular-flow economy based on increasing returns is constructed which yields unemployment equilibria as a natural consequence. Needless to say, constructing this kind of specialised model is more an exercise in logical consistency than a serious attempt to derive policy conclusions. While it is, ultimately, important to understand the microeconomic foundations of macro-theory and to reconcile classical and unemployment versions of general equilibrium theory, the benefits to a practising macroeconomist are likely to be quite indirect.[1]

II. INTRODUCTION TO THE MODEL: TASTES AND TECHNOLOGY

To focus as sharply as possible on basic issues, the paper deals with a very simple model having extreme symmetry properties.

Consumers derive their ultimate satisfaction from characteristics which are embodied in specific commodities. People differ in their preferences for the underlying characteristics. A person of a given attribute preference type wants commodities to be as close as possible to his type. Suppose there is a natural one dimensional preference ordering for attributes from O to H which allows them to be meaningfully represented as points on a circle of circumference H that begins at O and ends at H, where O and H coincide.

The population of N people is uniformly distributed in attribute preference types around the circle. Suppose the marginal rate of substitution between 'quantity' and 'quality' is constant. If a person of attribute preference type i is given x units of commodity j, his utility is

$$V_i(x,j) = x-\mu|i-j|, \tag{1}$$

where μ is a positive constant and $|i-j|$ represents the arc distance on the circle from i to j.

A consequence of the above formulation is that each person specialises in

[1] There is by now a vast literature on the microeconomic foundations of unemployment. It would be a tedious exercise to go through the various models *seriatim*, listing particular differences and similarities with the model of this paper. Instead, some general comments are offered. In contemporary models, the crucial role of Say's Law as an adjustment mechanism does not seem to be addressed directly. The literature does not emphasise the connection between macroeconomic unemployment and microeconomic barriers to entry. If there is monopolistic competition, the firm typically faces a 'perceived' or 'conjectural' demand curve rather than the 'real' demand of the present approach. An exception is the model of Hart (1982), which differs from the present formulation in a number of ways, including non-increasing returns, a fixed number of firms, and voluntary unemployment essentially caused by a 'too high' monopoly price. Although his search theoretic framework is quite different, Diamond's (1982) model served as an intellectual inspiration because of the way he grounds steady state involuntary unemployment firmly in basic principles.

the consumption of just one type of product. An alternative interpretation avoids this unrealistic feature while yielding the same aggregate behaviour. Think of each person as a composite of random preference atoms distributed uniformly around the attribute circle. Every atom is itself like a hypothetical modular consumer with the extreme preferences given by (1). The demand attributed to each person is the expected demand over uniformly distributed preference atoms. This interpretation, which makes everyone a generalist in consumption, is more like the formulation I have in mind, although the story is somewhat easier to set out in the analytically equivalent world where consumers are specialists.

To understand the role of differing assumptions about technology clearly, we resort to a fictional account of the evolution of a hypothetical economy which passes through three stages of technological development. Even though it is stretching history to talk about three modes of production as if they actually occurred in this precise sequence, it is nevertheless analytically very useful to imagine an economy that evolved this way.

In each case we are working with one factor of production. At the level of abstraction of an entire economy, this single factor is meant to represent a composite of every factor in the economy. Combining together all factors of production eliminates the possibility of *relative* factor unemployment and forces the model to focus sharply on the issue of *absolute* unemployment, which is the crux of an effective demand failure. We imagine, then, that various kinds of labour, capital, land, and natural resources can be miraculously aggregated into a meaningful index number, which we henceforth call 'labour' in honour of the major factor. Each of the N persons in our mythical economy is endowed with one unit of labour.

On the output side, no distinction is made between goods and services, between real and financial activities, or between production and distribution.

III. STAGE I: SELF SUFFICIENCY

Suppose each labourer can produce α units of any commodity. In such a world the economic problem has a trivial Robinson Crusoe solution. A person of attribute type i simply produces and consumes α units of commodity i. There is no need for trade in this economy, since everyone specialises in production of the commodity he most prefers. Of course it does no harm if we think of each person as hired to produce what his neighbour likes and then buying his own preferred commodity. Either way there is no possibility of unemployment because everyone has the fallback option of retreating into autarchy. With the price of a unit of each commodity fixed at one, the real wage in stage I is α.

IV. STAGE II: SMALL SCALE SPECIALISATION

Now suppose a person of type (i,j) prefers to consume i but has a comparative advantage in producing j. For simplicity, assume i and j are independently, uniformly distributed throughout the population. A person of type (i,j)

can produce β units of commodity j, or α units of any other commodity, where

$$\beta > \alpha. \tag{2}$$

Suppose a competitive market opens up where people can buy and sell commodities costlessly. The only possible competitive equilibrium occurs when the prices of all the goods are the same. In competitive equilibrium a person of type (i, j) produces β units of j and sells it in order to buy β units of i. If the price of commodity i is raised higher than the price of other commodities, the demand for i will fall below its supply, as some persons with attribute preferences near i switch over to close substitutes.

Although it is perhaps easiest to think of each person as self employed, there is no difficulty imagining a situation where several people with a comparative advantage in producing a certain commodity are hired for wages by a shadow firm. Under competitive equilibrium, the real wage w/p_i must equal β.

In such an economy there can be no true unemployment because there are no true firms. If anyone is declared 'unemployed' by a firm, he can just announce his own miniature firm, hire himself, and sell the product directly on a perfectly competitive market. While the 'unemployed' worker of type (i, j) is busy producing commodity j and selling it directly, an 'unemployed' worker of type (j, i) is simultaneously supplying the market with commodity i. The matching of unequal tastes is no problem because in such a balanced expansion supply creates its own demand.

Returns to Scale, Market Structure, and Say's Law: A Digression

The basic conclusions of last section extend to the modern formulations of classical general equilibrium theory. With hardly any loss of generality, the production side can be represented by a constant returns to scale activity analysis model. Diminishing returns is correctly viewed as a partial equilibrium phenomenon having to do with the effects on output of varying one set of inputs while another set is held constant.

Once granted the powerful assumptions of strict constant returns to scale and perfect competition, the essential logic of an adjustment mechanism which eliminates unemployment seems inescapable. Unemployment equilibrium is impossible in a constant returns world. To have a genuine theory of involuntary unemployment requires a genuine theory of the firm – i.e. an explanation of the organisation or process from which the unemployed are excluded.

In a constant returns economy the firm is an artificial entity.[1] It does not matter how the boundary of a firm is drawn or even if it is drawn at all. There is no operational distinction between the size of a firm and the number of firms. In a linearly homogeneous production system, it is immaterial which factor hires which. With sufficient divisibility, the unemployed factor unit can simply declare itself a miniature firm, hire itself and any other factors it needs, and sell the resulting output directly. Essentially the same idea works if there

[1] For a view of the concept of the firm implicit in classical general equilibrium theory, see, e.g., Koopmans (1957), Samuelson (1967), or McKenzie (1981).

is some indivisible scale of production small enough not to spoil the market. The operational requirement is that the efficient minimum-cost scale of production be sufficiently small, relative to the size of the market, that any one firm or plant cannot affect prices appreciably. The macroeconomic significance of perfect competition is that self-interested agents are motivated to eliminate unemployment *without* having to violate a 'natural' wage contour of equal pay for equal work.

When unemployed factor units are all going about their business spontaneously employing themselves or being employed, the economy will automatically break out of unemployment. While the simple story of supply creating its own demand can be told best in a closed barter economy, I do not see the existence of money, saving, investment, or international trade *per se* invalidating the basic proposition that a logical inference of strict constant returns to scale and perfect competition is full employment.[1] With sufficient divisibility in production, each unemployed factor unit has an incentive to produce itself out of unemployment and market the product directly. In effect, the unemployed are induced to create on their own scale an exact replica of the full employment economy from which they have been excluded.

Note that price flexibility guarantees (at most) the correct *relative* factor proportions in any general equilibrium production system with constant returns to scale. The *absolute* factor employment level represents an extra degree of freedom which must be pinned down by some direct quantity adjustment.

'Say's Law of Markets', the doctrine that supply creates its own demand, is understood here merely as a label for the kind of story being told about a quantity adjustment mechanism which increases output when there is slack capacity. The parable describes how an economy can automatically produce itself out of unemployment by a balanced expansion kind of bootstraps operation. For the purposes of this paper, Say's Law means that an exact scale replication, by the unemployed, of the production pattern of the employed economy will take place in a linearly homogeneous production system and that it is self-supporting because it generates an equi-proportionate increase in demand. However muddled its initial expositions, Say's Law represents an early piece of general equilibrium dynamics which is neither trivial nor unimportant.[2]

The role of Say's Law as an adjustment parable is crucial to the classical

[1] Drazen (1980) contains a careful rebuttal of the idea that money as a constraint on transactions is central to understanding unemployment. Typically, it is too much aggregation in simple models which leads to the mistaken notion that the cause of effective demand failures is the monetary exchange requirement. As for hoarding a non-produced asset, even with this feature the basic issue remains how, in a world of constant returns and perfect markets, equilibrium can persist when an unemployed factor unit could break itself out of unemployment by producing directly. While it is possible to tell more complicated and realistic stories, the financial system I have in mind for this essay uses just the simplest kind of pure inside money. Essentially money is credit. At negligible cost, banks issue short term credit to any firm that can repay loans for the hiring of productive factors. Note that the distinction between a money and a barter economy is blurred for the case of perfect competition, but the existence of a general unit of account in which factor payments must be denominated becomes important when producers are able to influence prices. The monopolistic firm faces a different maximisation problem with a different solution if, e.g., wage rates are set in kind rather than denominated in a general unit of account.

[2] This is essentially the interpretation of Schumpeter (1954).

belief that underlying forces tend to restore the economy away from 'temporary derangements' back toward full employment equilibrium. J. S. Mill was categoric about the importance of Say's Law: 'The point is fundamental; any difference of opinion on it involves radically different conceptions of Political Economy, especially in its practical aspect.' If Say's Law is not accepted, Mill wrote, economists must consider not only the allocation of a given volume of resources, but also 'how a market can be created for produce, or how production can be limited to the capabilities of the market'.[1]

Of course Say's Law of Markets has become a favourite whipping boy since the *General Theory*. But given classical assumptions of strict constant returns to scale and pure competition, (including symmetric information), the doctrine seems a perfectly logical theoretical proposition. Furthermore, although this is a separate issue, one suspects that Say's Law was probably not a terrible *empirical* description of how certain economies tended to adjust themselves automatically toward full employment at around the time Say was popularising the doctrine.

The world in which Say and Ricardo lived was an economic society that has since largely passed away and, somewhat ironically, was fading at the very time it served the classical economists as an inspiration for so many useful parables. This was a world where most producers were either peasant farmers or master craftsmen. When labour was hired, it was typically by individual proprietors on a small scale. Agriculture was the overwhelmingly predominant economic activity. To 'manufacture' – the word literally means 'make by hand' – was basically to employ oneself, perhaps with a few others, by setting up shop. The example most clearly epitomising this vanished world is the domestic system where cottage industry and agriculture were carried out by the entire family almost as an integrated enterprise.

Whether or not constant returns to scale and pure competition was ever a fair abstraction of a real economy, the Industrial Revolution represented an unequivocal triumph of the factory system with its overwhelming economies of scale. The reasons for increasing returns are anything that makes average productivity increase with scale – such as physical economies of area or volume, the internalisation of positive externalities, economising on information or transactions, use of inanimate power, division of labour, specialisation of roles or functions, standardisation of parts, the law of large numbers, access to financial capital, etc., etc. Constant returns means that average productivity is the same for all scales larger than some reasonably small indivisibility threshold. Only increasing returns causes a genuine separation of the aggregate factor from the product it produces.

This is not the place to get too involved in a debate about the actual extent of increasing returns in the real world. Economists who want badly enough to have a perfectly competitive economy will see increasing returns limited to a finite (and hopefully small – relative to the size of the market) initial production

[1] *Principles of Political Economy*, chapter XIV. Mill set out a more detailed analysis of the theory behind the classical position in his remarkable essay, originally written in 1829–1830, 'On the Influence of Consumption on Production'.

phase, after which the long run U-shaped average cost curves turn up, or at least become horizontal. For my part, I join others in seeing abundant evidence that significant economies of scale are a pervasive feature of the modern industrial world.

The seemingly institution-free or purely technological question of the extent of increasing returns is a loaded issue precisely because the existence of pure competition is at stake. The classical economists do not seem to face this issue head on, although their theories of value indicate they were often thinking in terms of constant returns to scale. Somewhat ironically, Adam Smith (whose invisible hand paradigm ultimately rests on convexity assumptions) also believed the division of labour is limited by the extent of the market, essentially a doctrine about the importance of increasing returns.

To emphasise a basic truth dramatically, let the case be overstated here. Increasing returns, understood in its broadest sense, is the natural enemy of pure competition and the primary cause of imperfect competition. (Leave aside such rarities as the monopoly ownership of a particular factor.) Once you agree you are observing a species of what might be called 'market economy', 'private enterprise', or 'laissez-faire', you have, so to speak, only two further sub-species to investigate. You are looking at pure competition if there are universal constant returns to scale in *every* aspect of technology (including the acquisition of knowledge and the transmission of information). You are looking at imperfect competition if there are increasing returns to scale.

Therefore, if you want to build from first principles a broad based microeconomic foundation to a general equilibrium theory that will explain involuntary unemployment, you must start with increasing returns and go the route of imperfect competition. Otherwise, you will forever be struggling in one way or another to evade the basic truth of Say's Law under strict constant returns to scale. Modelling the failure of coordination implicit in an 'inability to communicate effective demand' requires increasing returns and product diversity. With enough ingenuity, unemployment can be generated in models by various forms of asymmetric information like adverse selection or moral hazard, but division of labour strikes me as a rather more substantial foundation.

Of course the practical macroeconomist may wish to knock down the straw man of Say's Law at the very beginning and go right to a macroeconomic formulation of the problem. The task here is a different one. Instead of avoiding the issue by asserting, quite correctly, the empirical proposition that Say's Law does not hold in a modern industrial economy and moving on to practical matters, I want to sketch out what I think is involved in building up from first principles a consistent microeconomic foundation for unemployment equilibrium theory.

V. STAGE III: LARGE SCALE SPECIALISATION UNDER INCREASING RETURNS TO SCALE

Continuing with the metaphor of stages of technological development, suppose that increasing returns to scale in specialised production has been introduced.

More specifically, let the production function for any particular commodity be of the form

$$Y = \max [0, \gamma(L-F)], \tag{3}$$

where Y is output, and L is the total labour applied (the 'type' of labour does not matter in this mode), and γ and F are technological constants.

To produce any output at all in this new mode amount F of the factor (measured as a flow of services) must be committed to 'overhead'. It is intended that at 'reasonable' production scales, the average product of labour is significantly higher under large than under small scale specialisation,

$$\gamma\left(1-\frac{F}{L}\right) > \beta. \tag{4}$$

That way we can think of the new technology as clearly dominating the old.

In this paper the basic unit of production is the plant or the firm, no distinction being made between the two concepts.

The specific form of production function (3) is chosen because it is easy to work with and interpret. More general production functions showing increasing returns would give similar results.

At the aggregation level of the model, it is difficult to identify the source of the increasing returns (remember that L stands for all factors of production), nor does it really matter. Any or all of the usual reasons might be given. It is intended in this paper that the term 'increasing returns' be understood as a generic label for all phenomena causing average productivity to rise with scale.

For ease of exposition, we assume that the increasing returns mode represented by (3) is the only available method for producing commodities. An unemployed worker receives a zero wage instead of the comparatively miniscule amount to be earned by retreating into the 'second economy' of stage II or the even tinier reward available in stage I. We return to these matters afterwards.

Let the nominal wage paid by each large scale firm be parametrically fixed at some value w. As an expository simplification, suppose that aggregate demand is maintained at the hypothetical level which *would* be generated *if* the source of all spending were factor income at an unemployment rate u. We shall solve for the equilibrium prices and outputs of products and the equilibrium number of firms.[1]

Suppose there are m firms or plants producing m different products spaced equally around the attribute circle, where for the moment we think of m as some large exogenously fixed number. The distance in attribute space between the products of any two factor units is H/m. Let all plants charge a nominal price $\bar{p}$ for their product. We assume that firms set prices and react with quantities.

Allow any one firm to vary the price of its output. If the firm charges a price p, any worker specialised to buying the firm's product will purchase w/p units. In attribute space, suppose the firm is able to attract customers with

[1] The model is a general equilibrium version of monopolistic competition theory as developed by Hotelling (1929), Chamberlin (1933), Lancaster (1979), Salop (1979), and others. The underlying norm is the 'most perfect' competitive equilibrium that could possibly exist under increasing returns.

attribute types in a range of $-h(p)/2$ to $+h(p)/2$ centred on the particular commodity attribute the firm is producing. This is the firm's 'market area'. The marginal buyer must be located right on the boundary at an attribute distance $h(p)/2$ from the firm and $(H/m)-h(p)/2$ from the firm's nearest neighbour. By (1), the buyer who is just indifferent between purchasing from the firm or its nearest neighbour must satisfy

$$\frac{w}{p}-\mu\left[\frac{h(p)}{2}\right] = \frac{w}{\bar{p}}-\mu\left[\frac{H}{m}-\frac{h(p)}{2}\right], \tag{6}$$

which can be rewritten

$$h(p) = \frac{H}{m}+\frac{1}{\mu}\left(\frac{w}{p}-\frac{w}{\bar{p}}\right). \tag{7}$$

With u parameterising the level of short run aggregate demand, the total number of customers buying from the firm is $n(p)$, where

$$\frac{n(p)}{h(p)} = \frac{N(1-u)}{H}. \tag{8}$$

(The unemployed are uniformly distributed along the circumference of the circle.)

Combining (7) and (8),

$$n(p) = \frac{N(1-u)}{H}\left[\frac{H}{m}+\frac{1}{\mu}\left(\frac{w}{p}-\frac{w}{\bar{p}}\right)\right]. \tag{9}$$

With each customer buying w/p units, the total demand faced by the firm is

$$d(p) = \frac{w}{p}\,n(p). \tag{10}$$

Substituting from (9) into (10) and rearranging, the demand curve for the firm's product is

$$d(p) = \frac{N(1-u)\,w}{m}\left(\frac{1}{p}\right)+\frac{N(1-u)\,w^2}{H\mu}\left(\frac{1}{p}\right)\left(\frac{1}{p}-\frac{1}{\bar{p}}\right). \tag{11}$$

It will be convenient to introduce new constants a and b, and rewrite (11) as

$$d(p) = a\left(\frac{1}{p}\right)+b\left(\frac{1}{p}\right)\left(\frac{1}{p}-\frac{1}{\bar{p}}\right), \tag{12}$$

where

$$a = \frac{(1-u)\,Nw}{m} \tag{13}$$

$$b = \frac{(1-u)\,Nw^2}{H\mu}. \tag{14}$$

In a symmetric Nash equilibrium in prices, $p = \bar{p}$.

The elasticity of the demand function (12) evaluated at $p = \bar{p}$ is

$$E = 1+\frac{b}{ap}. \tag{15}$$

The marginal cost of producing an extra unit of output is

$$\frac{w}{\gamma}.$$

Using the profit maximising monopoly pricing formula[1]

$$p\left(1-\frac{1}{E}\right)=\frac{w}{\gamma} \tag{16}$$

and substituting from (15), we have

$$p=\frac{wb}{\gamma b-wa}. \tag{17}$$

Substituting from (13) and (14), expression (17) becomes

$$p=\frac{w}{\gamma\left(1-\frac{H\mu}{\gamma m}\right)}. \tag{18}$$

In $p=\bar{p}$ equilibrium, from (12),

$$d=\frac{a}{p}. \tag{19}$$

Substituting from (13) and (18) into (19),

$$d=\frac{\gamma(1-u)N}{m}\left(1-\frac{H\mu}{\gamma m}\right). \tag{20}$$

In the short run, the model treats as exogenously fixed: aggregate demand, the number of firms, and the nominal wage. Endogenously determined by profit maximisation are: prices, outputs and employment. From (18), the short period price to wage ratio is independent of aggregate demand. The profit maximising short term reaction to aggregate demand shocks is a pure quantity adjustment, which creates volatile pro-cyclical fluctuations of productivity and profits.

Now think of a longer term where m can be treated as variable. If there is free entry and exit, and if it is costless to relocate, the number of factor units should adjust to yield zero pure profits in equilibrium. (Remember, all factors are subsumed in 'labour' – there is no genuine entrepreneurial ability *per se*.) Behind this solution concept is some complicated dynamic story about the lure of eventual profits encouraging a new firm to establish a foothold between existing firms and to wait out the adjustment period of price warfare stubbornly until neighbouring firms are convinced to move aside in the product spectrum. The story can only be defended as an approximation. Entry and exit are complicated phenomena, involving difficult game theoretic issues that defy neat analytic formulation. If the number of plant openings is roughly pro-

[1] It can be verified that marginal revenue is falling and marginal cost is constant, so the second order conditions are satisfied. Note that when the firm hires a worker, except with negligible probability there is no effect on its own demand. If the workers of a firm consume primarily what that firm produces, the failure of coordination vanishes.

portional to existing profits per plant, then in equilibrium there will be zero pure profits,[1] or:

$$pd = w\left(F+\frac{d}{\gamma}\right). \tag{21}$$

Expression (21) can also be interpreted as an equilibrium condition guaranteeing that the total amount of factor which the firms wish to hire at a certain level of aggregate demand will actually generate that level of aggregate demand when the factor income is spent.

Substituting from (18) and (20) into (21) and solving the resulting equation for the equilibrium number of firms, we have

$$m = \sqrt{\frac{(1-u)\,NH\mu}{\gamma F}}. \tag{22}$$

Then substitute (22) into (18) and (20), yielding

$$p = \frac{w}{\gamma\left[1-\sqrt{\frac{H\mu F}{(1-u)\,N\gamma}}\right]}, \tag{23}$$

$$d = \gamma F\left[\sqrt{\frac{(1-u)\,N\gamma}{H\mu F}}-1\right]. \tag{24}$$

Substituting into (15), the equilibrium elasticity of demand for a firm's product is

$$E = \sqrt{\frac{(1-u)\,N\gamma}{H\mu F}}. \tag{25}$$

What follows now is a treatment of some technical issues. A more substantive discussion of the meaning of unemployment equilibrium is reserved for the next section.

While equations (22)–(25) have been derived as if u and w were given, in fact it is better to think in terms of simultaneous equilibrium relations holding among all the variables. When envisioning how Say's Law might operate under increasing returns, it is more useful to invert (22) and see m as determining u by the equation

$$u = 1-\frac{\gamma Fm^2}{NH\mu}. \tag{26}$$

If only there were an incentive to increase m, it would translate into an automatic mechanism for pushing down u.

[1] The story is somewhat easier to accept when it is borne in mind that each firm actually has many more neighbours than two, so that the impacts of changes in one firm are diffused over many adjacent firms. For example, in two dimensional preference space the analogous model would have every firm surrounded by a hexagonal array of six neighbours, instead of two in the analytically more tractable case considered here. There are other ways of closing the model than a zero profit condition, but they typically involve an implicit limitation on 'entrepreneurial ability' or whatever else is thought to lie behind an arbitrarily fixed number of firms. Such specifications are incomplete, with unemployment representing a comparative underutilisation of the tangible relative to the intangible factor rather than the absolute underemployment of all factors characteristic of ineffective demand. Closing the model by specifying any fixed rate of profit on sales would yield essentially similar properties to those derived here.

Equilibrium in the present context means that *real* magnitudes are constant, whereas no restriction is placed on the behaviour of nominal prices. At least in principle there could be proportional inflation or deflation.[1]

If equilibrium is defined to mean exclusively the Walrasian concept, there can be no involuntary unemployment in equilibrium by definition. It is, of course, possible to quarrel with usage of the word 'equilibrium' in this paper, but only semantic differences are involved so long as the substantive issues are understood.

Perhaps the time period most compatible with a variable number of factor units playing a significant adjustment role is some conveniently vague interpretation of the 'medium term'. Which equilibrium the economy converges to depends in general on dynamic specifications, adjustment speeds, and initial conditions.

From (22), (24), an increase in $N(1-u)$ causes both the total number of plants *and* the output of each plant to go up.[2] Deepening the extent of the market encourages firms to take greater advantage of economies of scale. This is an example of Adam Smith's famous doctrine that the division of labour is limited by the extent of the market. The case of interest is $N(1-u)$ relatively large, so that m is also big. This justifies the assumption that when a firm hires a worker there is no effect on its own demand. A large population also guarantees

$$FH\mu < (1-u)N\gamma, \tag{27}$$

a viability condition needed to ensure that monopolistic competition is sufficiently competitive to not degenerate into a form of monopoly.

We have been assuming that wages of the unemployed are zero. Actually, workers not employed in large scale industry are free to revert to the dual economy of Stage II. Provided m is sufficiently big to make the large scale firms relatively dense along the attribute circle, we can think if we wish in terms of two almost decomposable economies. The primary economy consists of large scale firms. The secondary economy is based on small scale specialisation of stage II. No one would commit to the secondary economy while expecting to earn more from searching for a job in the primary economy. Such a condition might be expressed in a form like

$$(1-u)\frac{w}{p} > \beta. \tag{28}$$

Equivalently, an unemployment rate u is viable whenever

$$0 \leqslant u < \bar{u}$$

[1] The model does not deal with the issue of how the nominal price level is determined, nor is it clear how best to 'overlay' such a model with inflation or deflation. Note that in principle it is easy enough to tell the same story about the real part of the model but allow wages and prices to change proportionately in accordance with some *ad hoc* version of an expectations augmented Phillips curve.

[2] Actually, the major adjustment is by varying the output of each existing firm. The fraction of total output change attributed to entry or exit varies from zero with high fixed costs to one half with low fixed costs.

for some upper bound value of unemployment $\bar{u}$ defined as satisfying

$$(1-\bar{u})\gamma\left[1-\sqrt{\frac{H\mu F}{(1-\bar{u})N\gamma}}\right]=\beta.$$

(This condition comes from substituting (23) into (28) holding with equality.)

Note that $\bar{u}$ is increasing in γ and decreasing in F, with

$$\lim_{\substack{F\to 0\\ \gamma\to\beta}}\bar{u}=0. \tag{29}$$

Condition (29) can be interpreted as saying that when a stage III economy comes close to being a stage II economy, unemployment must vanish in the limit. The reader can easily verify that as $F\to 0$ in (22)–(25), the economy approaches a perfectly competitive equilibrium: $m\to\infty$, $p\to w/\gamma$, $d\to 0$, $E\to\infty$.

Throughout this paper we are implicitly assuming that large scale specialisation is sufficiently more productive than small scale specialisation to make the ceiling unemployment rate $\bar{u}$ big relative to the actual unemployment rate u.

VI. UNEMPLOYMENT EQUILIBRIA

In the present model, *any* level of steady state unemployment is consistent with a self-fulfilling rational expectations equilibrium. The set of possible equilibria forms a continuum. When there is greater unemployment, purchasing power is diminished, which sustains a lower level of employment. If for any reason the economy gets stuck in a low employment equilibrium, left alone it will tend to remain there. The basic steady state tendency of the model can be fairly described as unemployment inertia.

From (23), the equilibrium real wage is

$$\frac{w}{p}=\gamma\left[1-\sqrt{\frac{H\mu F}{(1-u)N\gamma}}\right]. \tag{30}$$

If the nominal wage were changed in a kind of *ceteris paribus* experiment, the first order effect is a proportional change in aggregate demand. Consistent with the spirit of cost plus monopoly pricing, the nominal price of commodities should adjust in approximately the same proportion to leave the real wage and the unemployment rate unaltered. Implicit in such a statement is a dynamic adjustment mechanism which has the firms wait to see the new pattern of demand before revising price, output, and employment decisions. Ignoring indirect effects which are not formally included in the model,[1] the new profit maximising equilibrium would involve an equiproportionate change in prices, all real variables remaining the same.

An important inference of the present analysis is the idea that, in an increasing returns system, the equilibrium tradeoff between real wages and employment will tend to make ordinary wage adjustment mechanisms ineffective or unstable. Even if it could be done, an all round reduction of real wages cannot

[1] I am leaving aside redistributions of wealth like the real balance effect.

cure unemployment. Firms would find it cheaper to hire labour, but this effect is outweighed by a simultaneous decline in the demand for their products. Equation (30) shows that a successful attempt to depress real wages would actually *increase* the equilibrium level of unemployment.[1] The implication would seem to be that aggregate wage and price flexibility cannot make this kind of economy self correcting. Under such circumstances, wage stickiness may actually be a blessing.

The discussion in terms of hypothetical wage cuts is somewhat artificial. Perhaps a better way of stating the relation implicit in (30) is to say that equilibrium changes in employment are accompanied by procyclical movements in productivity and real wages.

Note that the potential for unemployment equilibrium is created by increasing returns in the production of a large number of different products. These are precisely the conditions of industrial organisation conducive to monopolistic competition. There is a sense, therefore, in which the natural habitat of effective demand macroeconomics is a monopolistically competitive microeconomy. Analogously, perfect competition and classical macroeconomics are natural counterparts.

Behind a mathematical veneer, the arguments used in the new classical macroeconomics to discredit steady state involuntary unemployment are implicitly based on some version or other of Say's Law.[2] It is true that under strict constant returns to scale and perfect competition, Say's Law will operate to ensure that involuntary unemployment is automatically eliminated by the self interested actions of economic agents. Each existing or potential firm knows that irrespective of what the other firms do it cannot glut its own market by unilaterally expanding production, hence a balanced expansion of the entire underemployed economy in fact takes place. But increasing returns prevents supply from creating its own demand because the unemployed workers are essentially blocked from producing.[3] Either the existing firms will not hire them given the current state of demand, or, even if a group of unemployed workers can be coalesced effectively into a discrete lump of new supply, it will spoil the market price before ever giving Say's Law a chance to start operating. When each firm is afraid of glutting its own local market by unilaterally increasing output, the economy can get trapped in a low level equilibrium simply because there is insufficient pressure for the balanced simultaneous expansion of all markets. Correcting this 'externality', if that is

[1] It is interesting to note that Keynes, writing after the *General Theory*, offered an explanation of procyclical real wages based on 'imperfect competition in the modern quasi-competitive system' where 'it is, beyond doubt, the practical assumption of the producer that his price policy ought to be influenced by the fact that he is normally operating subject to decreasing average cost'. See Keynes (1939), section V.

[2] To paraphrase Keynes, contemporary economists who might hesitate to agree with Say's Law do not hesitate to accept conclusions which require the doctrine as their premise. Rational expectationist 'unemployment by misperception' is a very different phenomenon from 'ineffective demand unemployment'.

[3] Note that any condition which 'blocks' Say's Law can cause unemployment to persist; increasing returns is merely the most convenient label to apply at a high level of aggregation. For example, imperfect capital markets can make it difficult to 'produce for oneself'; in effect, though, there is an increasing net return to ownership of the means of production.

how it is viewed, requires nothing less than economy-wide coordination or stimulation. The usual invisible hand stories about the corrective powers of arbitrage do not apply to effective demand failures of the type considered here.

In increasing returns equilibrium, at every unemployment level there is no incentive for a firm to hire more workers. Reducing wages in any one particular firm will certainly induce that particular firm to employ more workers because costs are reduced while demand is unaffected. But in a symmetric situation, wage pressure on *any* firm should represent wage pressure on *all* firms. If one firm is making a mutually profitable deal with the unemployed to work at a lower wage, so are other firms. At least as an approximation or norm, a law of one wage should prevail.[1] As we have seen, an all round change in the going nominal wage feeds back through aggregate demand and has no real effect, whereas cutting the real wage actually increases equilibrium unemployment.

With economies of scale, in zero-profit equilibrium there is a kind of natural barrier to further entry. If a new firm were to invade an existing market area, its size would be too large not to depress the prices it and its neighbours receive before adjustments can occur in the product spectrum. An additional lump of unbalanced supply will spoil the existing market before getting a chance to create its own demand. When there are no pure profits to begin with in such situations, Say's Law is frustrated from operating effectively. Under anything resembling a uniform wage structure, no firm would have an incentive to enter because it must suffer an initial discrete loss followed, at best, by zero profits in the long run. Who would underwrite market penetration experiments of that sort? If, for whatever reason, the law of one wage paid by all firms to all employed workers is accepted as a valid approximation, then only in the limit as the degree of increasing returns is negligible can there be intentional entry into a zero profit market.

Economists looking to blame unemployment on wage inflexibility err in thinking that *aggregate* wage stickiness *per se* has anything much to do with the matter. The crucial issue is whether or not there can be sufficient deviation from a *relative* wage contour of equal pay for equal work to overcome the barrier of increasing returns – an unlikely, asymmetric, socially disruptive adjustment mechanism, the need for which never arises under constant returns and perfect competition. The classical approach, after all, promises that involuntary unemployment can be eliminated if only *overall* wage levels are flexible, without requiring significant wage discrimination between otherwise

[1] Although it represents a simplifying symmetry postulate which, in the spirit of macroeconomic theory, allows us to speak of *the* aggregate wage level, strictly speaking a law of one wage paid by *all* (existing and potential) large scale firms to *all* (previously and newly) employed workers is an exogenous specification when the labour market is not clearing. In a Walrasian equilibrium, competition would drive all wages toward equality as well as, more importantly, establishing an aggregate wage level consistent with full employment. In the present model competitive pressure is allowed to influence only the aggregate wage level, without altering its relative profile. Behind this notion is a crude symmetry assumption – in the specification of the underlying game, wage pressure on all (existing and potential) firms is more or less identical. At the very minimum, this kind of assumption can serve as a point of departure. The idea that full employment should hinge so crucially on breaking a *relative* wage contour of equal pay for equal work is entirely alien to the classical tradition.

identical factor units differing only in whether or not they happened to be employed last period.

At any equilibrium employment level of the present model, total spending on the product equals aggregate factor payments. This reflects the idea that the primary source of current purchasing power is current income. While *systematic* deviation from this steady state norm is not to be expected, trade can become uncoupled in a monetary economy and intended aggregate demand for commodities may instantaneously exceed or fall short of factor cost for a variety of reasons. Indeed, volatility of aggregate demand is a basic theme of modern macroeconomics. When they occur, such autonomous spending shocks are disequilibrating. The amount demanded is more or is less than what has been produced. Factor units will be induced to expand or to contract in size, and then in number as a reaction to pure profits temporarily going positive or negative. Without further shocks, the system will gravitate toward some new equilibrium state at higher or at lower employment levels, where it will tend to remain. Naturally, the exact details of any particular dynamic adjustment depend on specific assumptions, but the broad tendencies should be reasonably clear.[1]

If intended aggregate spending can shift over time but within any adjustment period bears a sufficiently stable and systematic short term relation with national income to be considered a true function of it, then equilibrium is reestablished at the 'Keynesian cross' where the two are equal, and the usual multiplier effect can take hold.

Hysteresis effects are a significant part of the unemployment story being told here.[2] Although the model undoubtedly exaggerates the phenomenon of persistence, basically the system is sticky and tends to remain where it is unless there are external disturbances which change the state variables.

In some sense the fundamental message is that if for any reason a recession becomes convincingly protracted – whether the original cause lies in stock market crashes, oil price increases, curtailed investment spending, or whatever other shocks – an automatic mechanism such as Say's Law will not necessarily operate to draw the economy back to its previous level homeostatically. Like a self-fulfilling prophecy, once the expectation of aggregate demand associated with a given employment level becomes ingrained, it turns into a major cause of its own persistence. If plants are closed and business confidence is damaged, the condition tends to replicate itself. That sort of effect was almost surely the main source of the momentum that kept the Great

[1] Of course, if the economy is continually peppered with shocks it may never attain steady state equilibrium. Even in the most extreme cases of semi-permanent disequilibrium, I would argue it is important to understand whether the endogenous equilibrating tendencies are toward full employment or not. This paper omits an explicit dynamic analysis because preliminary results are messy and do not appear to be especially informative.

[2] *The American Heritage Dictionary of the English Language* defines hysteresis as 'the failure of a property that has been changed by an external agent to return to its original value when the cause of the change has been removed'. Just because the present model has Markovian features, it would be misleading to say the unemployment level is indeterminate. The present equilibrium state of the system is a cumulative response to past shocks, all effects of which are completely summarised in the present equilibrium state of the system.

Depression going, even though economists to this day are unsure what started it in the first place.

The model is much too crude for specific policy prescriptions, but it does suggest a pump-priming government strategy aimed at shocking a depressed economy into states of full employment and keeping the pump running long enough to build up perceptions of high aggregate demand. If total spending can be maintained at a level consistent with some hypothetical employment rate, at least in principle the employment will actually materialise and become self sustaining.[1]

VII. CONCLUSION

Is involuntary unemployment possible in a steady state equilibrium where intended aggregate spending equals total factor income?

The 'classical' answer is no. In one form or another, Say's Law will cause an economy to automatically break out of unemployment.

The 'Keynesian' answer is that underemployment can persist because self correcting forces are weak or non-existent. Grounding such an idea firmly in basic principles has been a major challenge to economic theory.

This paper argues that increasing returns to scale, understood in a broad sense, is the primary obstacle blocking unemployed factor units from producing on their own. When Say's Law is thus stymied, the easy road to automatic full employment is closed.

Massachusetts Institute of Technology
Date of receipt of final typescript: April 1982

References

Chamberlin, E. H. (1933). *The Theory of Monopolistic Competition.* Harvard University Press, first edition 1933, eighth edition 1962.

Diamond, P. A. (1982). 'Aggregate demand management in search equilibrium.' *Journal of Political Economy*, forthcoming.

Drazen, A. (1980). 'Recent developments in macroeconomic disequilibrium theory.' *Econometrica*, vol. 48 (2), pp. 283–306.

Hart, O. D. (1982). 'A model of imperfect competition with Keynesian features.' *Quarterly Journal of Economics*, no. 386, pp. 109–38.

Hotelling, H. (1929). 'Stability in competition.' Economic Journal, vol. 39, pp. 41–57.

Keynes, J. M. (1939). 'Relative movements of real wages and output.' Economic Journal, vol. 49 (193), pp. 34–51.

Koopmans, T. C. (1957). *Three Essays on the State of Economic Science.* New York: McGraw Hill.

Lancaster, K. (1979). *Variety, Equity, and Efficiency.* New York: Columbia University Press.

McKenzie, L. W. (1981). 'The classical theorem on existence of competitive equilibrium.' *Econometrica*, vol. 49 (4), pp. 819–42.

Salop, S. C. (1979). 'Monopolistic competition with outside goods.' *Bell Journal of Economics*, vol. 10, pp. 141–56.

Samuelson, P. A. (1967). 'The monopolistic competition revolution,' in Kuenne, ed., *Monopolistic Competition Theory: Studies in Impact*, Wiley.

Schumpeter, J. A. (1954). *History of Economic Analysis.* New York: Oxford University Press.

[1] The model hints at one intriguing policy implication which can only be briefly sketched here. An important part of the reason why an economy can get stuck in a low level equilibrium trap is that the wage of each firm is tied to a general unit of account over which the firm has no control. Consider what would happen if instead of a money wage, each firm paid a wage expressed in terms of its own product, or its profit per worker, or its revenue per worker, or any other firm specific currency which declines in value as more labour is hired. Now when each firm seeks to maximise profits there is an inherently expansionary bias and the only possible rest state is at full employment. The numeraire in which factor payments are denominated can very much matter when producers are able to influence prices, and a properly designed wage system can exploit this feature to avoid unemployment. These ideas will be pursued in a later paper.