Monopsony as a metaphor for the emerging post-union labour market

Christopher L. ERICKSON* and Daniel J.B. MITCHELL**

Abstract: How can employers worldwide be experiencing increasingly severe labour shortages in the face of globalization? Why don’t wages rise in expanding economies? This article argues that declining union power has allowed employers to take the upper hand, setting pay and other conditions of employment as they would in a monopsonistic labour market. Rejecting the perfect competition model matching supply to demand, the authors argue that, far from being a pedagogical curiosity, monopsony’s imbalance in bargaining power is widespread. Employee voice needs to be restored to counter the undesirable consequences of strong macroeconomic performance, such as wage inequality and reduced worker rights.

In essence, this article deals with the following puzzle. How is it possible that, in the face of apparent ongoing globalization of product, capital and labour markets, employers everywhere, throughout the world and across industries, seem to be facing increasingly severe labour shortages? For example, the magazine Business Week, one of the most prominent voices of American capitalism, recently ran an article entitled “Where are all the workers? Companies worldwide are suddenly scrambling to manage a labor crunch”. The authors of the article reported that employers in such far-flung locations as Viet Nam, Bulgaria and the United States claim to be unable to find people “who are both able and willing to do the work for the posted pay” (Coy and Ewing, 2007). Our argument is that the answer to this puzzle – and to several others in the current economy, such as the declining share of income going to labour in the face of rising productivity – is the combination of union decline and an overlooked institutional feature of the labour market, monopsony.

* Professor, UCLA Anderson School of Management. Email: chris.erickson@anderson.ucla.edu. ** Ho-su Wu Professor, UCLA Anderson School of Management and UCLA School of Public Affairs. Email: daniel.j.b.mitchell@anderson.ucla.edu. We thank Sanford Jacoby for helpful comments on earlier drafts.

Responsibility for opinions expressed in signed articles rests solely with their authors and publication does not constitute an endorsement by the ILO.

Copyright © International Labour Organization 2007
Background on competition and monopsony

Observers of wage and employment dynamics have long viewed monopsony power as a credible, if problematic, phenomenon in “free” labour markets (see, for example, Bronfenbrenner, 1956). Recently, at least partly in response to the baffling economic conditions of the 1990s, some observers have argued that monopsony (or oligopsony) conditions characterize a wider range of particular micro labour markets than was previously believed (Bhaskar, Manning and To, 2002; Card and Krueger, 1997; Manning, 2003).

We go beyond the analysis of particular labour markets here to argue that monopsony serves as a powerful metaphor for the emerging post-union macro labour market, particularly in the United States but also, by extension, throughout the world. Wherever unions are losing the power to influence wage setting meaningfully – or where they have never had the power to do so – monopsonistic setting of pay and other conditions of work exists. We will present a variety of evidence that, taken together, suggests that the labour market, in the absence of unions, is characterized by employer bargaining power.

In the perfect competition model, wage equals marginal revenue product of labour. But under monopsony, wages are below marginal revenue product, i.e. income is shifted from labour to capital, resulting in a tendency toward labour shortages during “Good Times”. Among the most important implications of this argument is the fact that the recent diminution of union power does not, per se, lead to significantly higher employment rates, ceteris paribus, but rather, primarily, to lower wages and pervasive employer perceptions of labour shortage. In other words, there is a transfer of “rents” from labour to capital. While there are certain benefits of a monopsonistic labour market at the macro level, there are also micro-level costs.

In a sense, the question we are addressing is the mirror-image of the classic puzzle of the Great Depression: the question is not the well-known “Why don’t wages fall in a contraction?” (Bewley, 1999), but rather “Why don’t wages rise sufficiently to clear the market in an expansion?” Our answer is that the decline of unions and other sources of worker power have allowed employers to gain the upper hand and to act as they would in a monopsonistic or oligopsonistic micro labour market. We will present evidence at the economy-wide level in the United States which could be indicative of monopsony, and a simple model of the macroeconomy showing the dynamics of employment and wages under the monopsonistic assumption. We will also explore possible alternative explanations for the recent developments in macro labour market outcomes and compare them with monopsony.

Recent developments in labour markets

The 1990s were a puzzling period for economists and other observers of labour markets in the United States. Indeed, the period was sufficiently surprising that a substantial literature developed to explain it (Krueger and Solow, 2001).
The economy of the United States surged by measure of any indicator, particularly during the second half of the 1990s. And yet inflation – whether measured by prices or wages – remained surprisingly low despite a fall in unemployment to levels not seen since the Vietnam War. Accompanying this drop in unemployment was a reduced use of lay-offs, a conjunction which, by itself, is not especially surprising. A surprising feature, however, is the correlation between fewer lay-offs and proliferating labour shortages, we will argue.

Unfortunately, there is no ongoing time series on lay-offs for the United States labour market. A proxy is the rate of new claims for unemployment insurance (UI). As figure 1 shows, not only did that figure fall in the late 1990s, it did so to below Viet Nam-War levels. Moreover, at recession peaks (designated by the arrows on the figure), lay-offs in the recessions of the early 1990s and of 2001 ran at substantially lower rates than in earlier recessions.

Other notable changes in labour market outcomes that have now been widely accepted as facts include the widening of earnings inequality throughout the world during the last few decades (e.g. Katz and Murphy, 1992; Juhn, Murphy and Pierce, 1993; Acemoglu, 2003), and the steady decline in labour’s share of national income in the United States since the early 1980s.\(^1\)

Finally, two additional key characteristics of the “new economy” are a drop in labour force participation and widespread labour shortages. Research at the Federal Reserve Bank of San Francisco (Motley, 1996), the Federal Reserve Bank of Atlanta (Hotchkiss, 2004), and the Federal Reserve Bank of Minnesota (McMurry, 2002) has highlighted the decline in labour force participation (particularly among prime-aged males) and its plausible effect on the economy and on growth in particular. The overall labour force participation rate – which rose steadily after the Second World War owing to increased female participation – peaked in the late 1990s and then declined. Female participation also stopped rising after 2000 and, as late as 2006, was below its peak level.

As for labour shortages, it is first important to acknowledge the inherent difficulty in ascertaining an increase or decrease in vacancies, or labour shortages. Counts of newspaper “help wanted” advertisements were once viewed as reliable indicators, but have lost meaning with the rise of new technology, such as on-line job-search services. We present statistics of newspaper articles on Nexis/Lexis mentioning labour shortages in recent years. Figure 2 indicates that the overall number of articles on shortages increased dramatically in the late 1990s. Figure 3 indicates that these shortages were not limited to one or two sectors such as high technology, but were in fact mentioned in a variety of industries with varying skill requirements.

---

\(^1\) Labour’s share of corporate income began to fall in the 1980s, reversing a long-standing rising trend. The share of corporate national income going to labour is highly cyclical. Figures for the postwar period at rough business cycle peaks are: in 1949, 76.6 per cent; in 1959, 78.1 per cent; in 1969, 80.3 per cent; in 1979, 82.3 per cent; in 1989, 81.8 per cent; in 2000, 79.6 per cent.
Figure 1. New weekly claims for state unemployment insurance (UI) as percentage of covered employment, 1946-2006

Per cent

Note: Arrows represent troughs in the business cycle.
Monopsony as a metaphor for the emerging post-union labour market

Figure 2: Labour shortage citations in United States sources, 1985-2003
Figure 3. Labour shortage citations in United States sources, by sector, 1985-2003.
A sampling of the articles themselves indicates shortages in lower-skill occupations across a variety of geographical areas: construction workers in Maryland, restaurant workers in Colorado and a general labour shortage in Wisconsin, just to name a few in addition to the international evidence compiled in Coy and Ewing, 2007. If one is to believe the press, growing labour shortages occurred well beyond particular occupations and places, such as computer programmers in Silicon Valley. It seems that, despite flat wage growth, especially at the bottom of the distribution, employers complained of widespread labour shortages.

In what follows, we offer an explanation for the surprising macro performance of the late 1990s, which was based on changes in labour market institutions and on what we term “employer ascendancy,” which took place largely to the detriment of workers’ bargaining power and wages. This shift has exacerbated one of the features that distinguished the United States from other developed economies: the creation of many jobs at low wages, compared to the “insider-outsider” European systems’ high wages and high unemployment rates (see, for example, Blanchard and Summers, 1988; Krugman, 1994). However, as labour markets deregulate and decentralize worldwide, it could become a global norm, and strong macroeconomic performance could continue and indeed spread. A number of undesirable social and economic consequences will nevertheless accompany the strong macroeconomic performance.

Union decline
The most dramatic change in the structure of the United States labour market since the 1980s has been the decline of unionization. This change was termed a “transformation” of industrial relations in a well-known book that appeared in the mid-1980s (Kochan, Katz and McKersie, 1986). The book focused primarily on the micro side of the story. But the pace of that transformation accelerated markedly in the 1980s and continues today. As the micro institutions of the labour market change, it is inconceivable that the transformation would not have macro impacts as well. Below we argue that, with regard to the macroeconomy, the transformation of industrial relations to a non-union regime of employer ascendency occurred largely in the 1980s. By the recession of the early 1990s, its impact was already detectable.

Union wage settlements in the 1960s and 1970s were a major target of macroeconomic policy through various wage-price guideposts and wage-price controls programmes. By the mid-1990s in contrast, the United States Bureau of Labor Statistics stopped keeping track of union settlements altogether.

---

2 See also Hussey (1999), Wagner (1999) and Johnson (1999).
3 See Waldinger and Erickson (2003) for a discussion of skilled labour shortages in the Silicon Valley.
4 Among the consequences are: lost employee voice, increased wage repression and income inequality, reduced enforcement of labour and employment laws, shifts from defined-benefit pension plans, reduced health insurance coverage and other job protections, etc.
Apparently, economic policy-makers in the Clinton era were no longer sufficiently concerned about such settlements for the cost of monitoring them to be warranted. Yet, despite lack of interest by official statisticians, and despite the tendency of researchers to overlook union decline as a source of economic change, the possibility that union erosion has had important macro consequences is worth considering.

Economists have often modelled unions as analogous to product market monopolies. Monopolists set their price above the competitive level and then sell the corresponding quantity. The academic literature has similarly tended to view collective bargaining as determining an above-market (monopolistic) wage, employers then autonomously setting employment, given that wage. In what follows, we take that approach when union wage determination is represented.

It might be assumed that the decline of unions should have made the United States labour market more “competitive.” The question is what being competitive implies in the context of a non-union labour market. Truly competitive markets are highly sensitive to shortages and surpluses and quickly eliminate both through price adjustments. Thus, the labour shortages of the last half of the 1990s in the United States labour market should have resulted in dramatic wage increases, if the markets had been highly competitive in the textbook sense.

The fact that there was no wage explosion in the late 1990s despite labour shortages, suggests that, if indeed labour markets are now more competitive, the characteristics of such “competitive” (i.e. non-union) markets need further elaboration. Labour markets clearly do not work the way competitive markets for treasury bonds or wheat futures do. We argue below that a reasonable model of such a market under employer ascendancy – the post-union labour market – is monopsony not perfect competition.

Monopsony/employer ascendancy in the labour market

Traditionally, labour-market monopsony models were confined to special cases such as company-coal mining towns or employer collusion in specialized occupations. The textbook case of a monopsony is one where there is a single employer in the labour market. Such an employer can only increase its work force by in-

5 It has been pointed out that such wage setting represents “inefficient” bargaining. The union should in theory bargain over both the wage and the employment level. However, the notion that the employer sets employment seems to correspond broadly to reality, with some notable provisos about workrules and staffing requirements, in certain cases.

6 The accelerated weakening of American labour unions since the early 1980s is a well-known story by now. Its most visible symptom was a substantial concession movement in which wages and benefits were frozen or cut, sometimes through unscheduled contract reopenings (Mitchell, 1994). Along with these concessions came a significant decline in union membership and coverage in private employment – a decline that actually switched from mere erosion relative to the workforce to absolute decline in the early 1980s. Moreover, the idea that weakened unions might lead to a lower non-accelerating inflation rate of unemployment (NAIRU) is also not new and goes back to the mid-1980s at least (Mitchell, 1986). In loose terms, anything that reduces “wage-pushiness” (or price-pushiness) will have that effect (Mitchell and Zaidi, 1992).
creasing the wage rate to all workers, in order to attract the marginal worker. With the marginal cost of labour (the cost of hiring the next worker) greater than the average cost of labour (the wage), the profit-maximizing monopsonist will stop hiring where the marginal revenue product (or the value of the output generated by the next worker hired) is therefore above the wage rate (see, for example, Flanagan, Smith and Ehrenberg, 1984, pp. 67–69). The bottom line is that, under a textbook monopsony, a single employer would pay a lower wage and hire fewer workers than would be the case if the labour market were competitive in the textbook sense. The employer would always face a shortage of workers, being willing to hire more employees at the going wage, but unwilling to raise the wage to attract more workers.

Monopsony can be thought of as the flip side of monopoly. Under monopoly, the seller has market power (whereas monopsony involves buyer power). A monopolistic firm restricts output and sells at a price above what would prevail if the product market were competitive in the textbook sense. Firms with monopoly power have a shortage of customers in that they would like to sell more if only it did not entail lowering the price. This shortage of customers accounts for such phenomena as advertising of branded products (which gives a measure of monopoly power to producers) and cheating cartel members (who sell more than their agreed quotas).

Monopsony was treated in labour economics textbooks as an exception to the assumption that unions inevitably faced a wage–employment trade-off in bargaining, i.e. a negatively-sloped demand curve. In monopsony, a counter-acting union monopoly effect could raise wages without necessarily decreasing employment, because in a monopsonistic labour market the wage is below the marginal revenue product of labour. However, the monopoly vs. monopsony idea was depicted typically as a curiosity, pedagogically useful in helping students learn the microeconomics of the labour market and, perhaps, helpful in explaining unionization of certain odd groups such as nurses and professional athletes.  

**Growing interest in labour-market monopsony**

During the 1990s, however, the monopsony (or oligopsony) model was applied more generally to low-wage labour markets by Card and Krueger (1997, pp. 355–386) as an explanation of the seemingly few jobs lost as a result of increased minimum wages. Although the empirical findings of Card and Krueger led to considerable professional and policy debate, labour-market monopsony thereafter began receiving more general attention in microeconomic analysis (Boal and Ransom, 1997; Manning, 2001). Bhaskar, Manning and To (2002,  

---

7 Freeman and Medoff (1984) put forward a model of unions with a “monopoly face” and a “voice face”. However, they do not focus on the notion that the monopoly side of unions might be a response to monopsony on the employer side. Implicitly, they assume that the alternative to union monopoly power is a labour market characterized by demand = supply set by classical competition.
p. 172), for example, argue that “the main sources of oligopsony power are likely to be preference heterogeneity over jobs, mobility costs and imperfect information.” Whatever the explanation, whether explicit collusion, implicit collusion via such instruments as wage surveys, the decline of union wage-standardization power, or the sort of preference matching and labour market frictions emphasized by Bhaskar, Manning and To (2002) and Manning (2003), there are clearly many potential sources of employer power to violate the “law of one wage”. Moreover, in our view, the monopsony model can usefully be extended to the macro arena to explain the puzzles that became evident as union representation declined.

In the treasury-bond market, to take an example, the product involved is homogeneous. Buyers do not need to evaluate the particular T-bonds they are buying; for a given issue, all T-bonds are the same. Sellers (buyers) do not care about the attributes of the buyer (seller) so that questions of “matching” do not arise; no ongoing relationship between buyer and seller needs to be established. The size of the T-bond market is such that no significant costs are incurred when searching for a buyer (seller). Finally, no issues of fairness constrain price movements in the impersonal T-bond market. Sellers of T-bonds may be disappointed if the price of what they are offering falls, but they do not believe that buyers have a moral obligation to maintain a particular price. Since they do not have an ongoing relationship with the buyers, even if they held such beliefs, sellers would not have the means to enforce them.

Labour-market attributes

The labour market is very different. Labour is not homogeneous. As both workers and jobs have unique characteristics, issues of matching clearly do arise. In most cases, some form of ongoing buyer-seller relationship is established. Job searches by workers and recruitment firms and employers’ screening costs can be considerable in terms of time and money. Because employment takes place in a social context, issues of fairness also arise. As will be explained more fully below, these characteristics of the market result in, from the employer’s point of view, an upward-sloping labour supply curve, regardless of the slope of the aggregate labour supply curve. Buyers who face upward-sloping supply curves are inherently monopsonistic. They set the wage they are willing to pay, based on that curve – wages are not set by some impersonal auction market.

Below we elaborate on three possible explanations of employer monopsony. The first is similar to the Card-Krueger approach. But other mechanisms repressing non-union wages can be at work, including patterns and tacit wage coordination by employers and, perhaps paradoxically, the phenomenon of downward nominal wage rigidity.

Labour flows and the labour supply curve

It does not take much analysis to generate an upward-sloping labour supply curve to the firm along Card-Krueger lines. A simple flow model suffices. Let
employment at a firm at time \( i \) be designated by \( L_i \). Let the number of workers hired in a given period (after recruitment and screening) be \( H \). Finally, let \( t \) represent the turnover (quit) rate during that period. Then \( L_2 = (H - tL_1) + L_1 \).

If the firm is in a steady state (with no employment change), then \( L_2 = L_1 \) and therefore \( H = tL \). The turnover rate is inversely correlated with the firm's wage. A smaller percentage of workers in any given period will quit if the wage is raised, since the opportunity cost of quitting is thereby increased. Figure 4 depicts the turnover rate function. For any given level of employment, the turnover rate function can be translated into the corresponding turnover number function by multiplying \( t \) by the level of employment, \( L \). Figure 5 depicts two such turnover number functions corresponding to employment levels \( L^* \) and \( L' \), where \( L^* > L' \).

Figure 6 illustrates a typical hiring function.\(^8\) The number of workers who are recruited will increase with the wage offered. Note that the turnover number function depends on the employment size of the firm while the hiring function is independent of firm size. (It effectively depends on the size of the much broader external labour market.)

The labour supply function perceived by the firm can be derived from the turnover number function and the hiring function. Consider figure 7 represents a situation where the wage set by the firm is such that the number hired in just equals the number quitting in a period.\(^9\) Let the corresponding steady-state wage and employment levels be denoted by \( W^* \) and \( L^* \) respectively. The steady-state wage and employment combination is found at the intersection of the \( H \) and \( T = tL^* \) functions.

Suppose now that the firm, for some reason, cuts the wage to \( W' \). Fewer workers will be hired by the firm per period and more workers will quit. Employment will begin to decline as outflow exceeds inflow. The turnover number function will begin to shift down and to the left. Even with a higher turnover rate, the number of quitting workers will tend to diminish over time as the base \( L \), to which it refers, falls. Eventually, a new steady state will be reached with employment at \( L' \), depicted in figure 7 as the intersection of the \( H \) function and \( T = tL' \).

Thus, a lower wage is associated with a lower level of firm employment, i.e. the labour supply curve, as seen by the firm, is upward sloping.\(^10\)

---

\(^8\) The intersection of the two axes of figures 6 and 7 is not at the (0,0) point. In particular, the wage at the intersection point is \( > 0 \) since \( H > 0 \).

\(^9\) See the previous footnote.

\(^10\) Although we do not pursue the matter here, it might be noted that the same model could be applied to product markets in which products are heterogeneous while matching and searching are issues. A firm might be depicted as having a customer base and drawing orders from that base. Raising prices will tend to erode the customer base, as fewer new customers are gained per period and a higher rate of attrition for old customers is experienced. Thus, high prices will lead to reduced demand and low prices to increased demand. The firm thus faces a downward-sloping demand curve, i.e. it operates in a form of imperfect competition. Note also that the rise of Internet technology might affect the \( H \) and \( T \) functions by reducing information costs: both \( H \) and \( T \) could become steeper as information costs diminish, thus flattening the labour supply curve and reducing monopsony power.
outflows (the vertical axis of figure 7) are also lower at the lower wage. This upward sloping labour supply curve is the essential condition for monopsony.

**Union wage determination, perfect competition, and monopsony**

Figure 8 presents basic labour market outcomes under three extreme regimes: pure union wage determination, perfect competition and monopsony. The labour supply curve perceived by the firm is the upward-sloping line $W$. Labour demand in the short run is the marginal revenue product of labour $M{RP}_L$ (representing the marginal contribution of each additional worker). Owing to the upward-sloping nature of the supply curve, the marginal cost of hiring is shown by line $MW$ which is greater than $W$ except on the vertical axis. This is because, in order to hire another worker, the assumption is that the employer must pay the higher wage rate going to this additional worker to all
of its existing workers as well (see, for example, Flanagan, Smith and Ehrenberg, 1984, pp. 67–68).

The perfectly competitive outcome would be at the intersection of the demand and supply curves, with equilibrium employment $L_p$ and wage $W_p$. The well-known union-determined outcome is at the intersection of the demand curve and $MW$, with employment level $L_u$ and wage $W_u$, with the union maximizing (and capturing) all of the available “rents.”

However, the firm’s optimal employment-wage decision is to set the wage at $W_m$, producing an employment level $L_m$. This is the same employment level as the union-determined outcome, but with the firm now capturing all of the “rents.” Note as well that at this employment-wage decision – the monopsony outcome – the firm has a labour shortage since the marginal revenue product of labour $W_u$ is greater than $W_m$. If the firm could somehow find a marginal worker
willing to work and stay in employment at $W_m$, it would make the added hire. But it will not raise the wage to attract such a worker, even though the obtainable marginal revenue exceeds the wage. The firm thus operates with the chronic labour shortage associated with monopsony (which characterized the late 1990s).

Employers in such situations may complain about labour shortages. But if they seek to deal with the shortages, they do not do it by raising pay. Commonly, they seek through the political channel to allow more immigration from areas where wages are lower; such immigration, if permitted, provides the added hires at the going monopsonistic wage. Some argue that the marginal workers from abroad are needed because they will do the work that residents will not. The idea that residents will not do the work because pay is too low is inevitably rejected.

If we conceived the post-union labour market as a shift to perfect competition, wages would drop and employment would rise (relative to pure union wage determination). However, a shift from pure union wage determination to pure monopsony implies an even larger drop in wages, possibly with no change in employment, and a shift of rents from workers to employers. This is our basic argument: monopsony, not perfect competition, is the best metaphor for the post-union macro labour market.

If, in fact, the post-union labour market were now behaving as a monopsony, the outcome would be consistent with many of the basic empirical facts described above: the decrease in labour’s share of national income, the drop in labour force participation, the widespread labour shortages, demands that outside workers (immigrants) be admitted to alleviate the shortages, and perhaps even the widening of inequality (if higher-wage occupations are less likely to be monopsony-like). To explore the apparently increased stability of the macroeconomy, we next introduce some simple dynamics.

**Firm behaviour in the face of demand declines**

Figure 9 again presents the standard monopsony model of the employer, with the monopsony equilibrium now labelled at employment level $L_1$, wage $W_1$ and marginal revenue product of labour $M_1$. Suppose the firm experiences a demand decline to $\text{MRPL'}$. If completely unconstrained, it will adjust to the lower demand by reducing its wage to $W_2$ and its employment level to $L_2$. Some of the drop in demand will be absorbed by the wage and some by employment.

However, the firm can be reasonably assumed to be constrained by some degree of nominal-wage rigidity of the type reported in the 1990s by Bewley (1999, pp. 208–209, 432–433) and by many others (Fehr and Tyran, 2001; Levine et al., 2002; Mitchell, 1993; Howitt, 2002, p. 130). Indeed, even if wages were more flexible before the Great Depression than after, empirical evidence suggests that nominal wage rigidity – or at least resistance – characterized that pre-union era (Bordo, Erceg and Evans, 2000). Nominal wage rigidity, a long-
standing labour-market characteristic, is an appropriate addition to the dynamic monopsony model. If the nominal wage cannot fall and remains at \( W_1 \), the employer must “eat” the demand decline. Wages do not absorb the decline (by assumption). And with the wage rigid, there is no incentive for the firm to reduce employment! Since \( M_3 > W_1 \), there remains a labour shortage, though it is smaller than before.\(^{11}\) Effectively, the firm has “laid off” vacancies rather than real employees. Wage rigidity acts as a monopolistic force, offsetting (some of) the firm’s monopsony power.

In short, monopsonistic firms will experience chronic labour shortages and have a cushion of vacancies to “lay off” in the face of declining demand. Eventually, however, if the decline in demand is sufficiently severe, they will

\(^{11}\) The degree of the labour shortage can be measured by the amount by which the marginal revenue product of labour exceeds the wage.
reduce employment levels (e.g. a more extreme demand decline to \( MRP_{L^2} \), which would drive the employment level to \( L_3 \)).

Note that, writ large, the model suggests that rather mild recessions can be expected in a labour market characterized by monopsony owing to the anti-lay-off cushion. Furthermore, if booms follow recessions, we should see mild recoveries as well (until a return to the monopsonistic equilibrium). The mild recessions of the early 1990s and 2000s – in contrast with the more severe previous recessions – are in keeping with this model.

**General demand declines and labour supply increases**

The analysis so far has considered a demand decline experienced by a single firm. Such declines could occur as the result of particular circumstances related to that firm’s product. If only that firm, or a small number of firms, were involved, overall labour supply conditions would not be much affected. However, a more general negative demand shock – a recession - could increase labour supply to the typical firm as well as decrease demand. Such a labour supply increase could occur if a significant number of firms were sufficiently adversely affected for actual lay-offs to occur.

The impact of a labour supply increase of displaced workers would be mitigated if some firms remained in the monopsony range and experienced continued labour shortages. Such firms would tend to absorb the displaced workers, helping to maintain the overall level of employment. In short, even if some firms were pushed into the lay-off range, a monopsony economy would tend to be stabilized by other firms. Of course, a very severe negative shock might tip almost all firms into lay-offs. But generally, only mild swings in unemployment could be expected.

Labour supply could also increase because of exogenous factors. Foreign immigration into the United States ran at high levels in the 1980s, the 1990s, and beyond. The 1990s also saw an increase in the supply of low-skilled labour, primarily women, owing to federal and state welfare-to-work reforms (Lerman and Ratcliffe, 2001). Despite the supply increase, the labour market absorbed the newcomers and yet continued to exhibit worker shortages. The monopsony model, with its chronic labour shortage, accords with that result.

**Contrast with a union labour market**

What if a union were to come into existence when the firm was operating at its initial \( W_1 \) and \( L_1 \) combination on figure 10 and bargained with the firm to raise pay – how would wage setting differ? Presumably, if it could, the union would raise the wage to \( M_1 \). There is no employment loss entailed in going to \( M_1 \), just a transfer of rents from employer to workers. Note, however, that nominal wage rigidity at \( M_4 \) or above has a dramatically different effect in the face of a demand decline than it does below \( M_4 \). At \( M_4 \) or above, there is no monopsony effect to offset. Thus, a demand decline at a fixed nominal wage translates into a job loss.
Consider, for example, a union that bargained wage $M_1$, the maximum wage that entails no job loss starting from the non-union wage $W_1$. If demand falls to $MRP_L^*$, employment declines to $L_U$, on figure 10 in the union wage case, in contrast to *no loss* of employment, in the non-union monopsony case. Nominal wage rigidity matters a great deal to employment dynamics in the event of negative demand shocks, depending on whether there is monopsony in the labour market or not.

**Patterns vs. coordination**

We have shown that a non-union employer will typically have monopsony power, once the dynamics of labour flows are considered. Monopsony does *not* depend on there being a single employer in the labour market, e.g. the old
coal-mining company town case of the textbooks. However, the inherent monopsony in non-union wage setting could be reinforced by employer coordination of wage policies, a de facto buyers’ cartel. The long-standing stories of monopsony in the nursing labour market have involved explicit coordination by health provider/employers in urban areas, for example.

Notions of wage imitation in the union sector (pattern bargaining) are a traditional fixture of the industrial relations literature (Ross, 1948; see also Erickson, 1996, for a discussion of change and continuity in patterns in the “post-concession” era). The idea that pay should be set in comparison with relevant groups has also long been held by arbitrators called in to settle “interest” disputes (Bernstein, 1954, pp. 51–71). It has also been found that setting wages through comparisons is widespread in the non-union sector.

In fact, formally or informally, a near-universal element of non-union wage setting involves finding out what someone else is paying for similar workers. The information may be gathered through trade associations, government surveys, or simply a phone call to the firm down the street (Bewley, 1999, pp. 92–95). “Benchmarking” is a common management practice for evaluation and decision-making of all internal policies, including pay. Even in the face of labour shortages, employers are always concerned about the bottom-line implications of paying higher wages than their product-market competitors.

The line between innocent information gathering and cartel-like collusion is a fine one. The existence of tacit agreements not to compete for labour has long been noted in the research literature (Myers and MacLaurin, 1943, pp. 40–43). It need not be the case that all firms pay the same wage – clearly they do not – or that all provide the same percentage of wage increase. As long as pay decisions at one firm influence pay decisions at others, a certain level of coordination is occurring.

In a labour market where unions represent a significant fraction of the workforce and where the threat of organization is real to non-union employers, union wage setting will have an influence beyond the bargaining units where it occurs. Until the 1970s, non-union firms reportedly watched union settlements “very carefully”, and made pay decisions based on their observations to avoid being unionized (Foulkes, 1980, p. 166). Under such conditions, firms – whether unionized or not – are likely to operate in the wage range at or above $M_4$, in figure 10, either because they are forced to do so through bargaining or because they think it prudent to do so as a defensive measure.

However, such labour markets can be “tipped” into the below $M_4$ monopsony range if the union sector declines sufficiently and the threat of new organizing recedes. Such decline and threat reduction characterized the 1980s and 1990s. Thus, it is plausible that in that period the United States labour market went from monopoly to monopsony wage determination, a shift that other countries are undergoing as union wage-setting power declines. Furthermore, countries that never had strong unions – or that ban, discourage or control and co-opt them (as in parts of Asia) – would also exhibit monopsonistic labour markets.
Rigidity down, rigidity up

There may be still other explanations of monopsonistic behaviour in the labour market, apart from the flow story and the coordination story. In the European context, in particular, it has been argued that legislation providing job security for incumbent workers actually reduces employment and hiring. The reason for this outcome is said to be risk avoidance. Employers know that if they hire workers and then need to reduce their workforce, terminations may be difficult, so they avoid risk by hiring fewer workers than they would under laissez-faire conditions (OECD, 1999, pp. 47–132). They can be said to “insure” themselves against the risk of expensive lay-offs by paying the costs of having too few workers (or using routes around the mandates such as temporary staff or off-the-books hiring).

The United States does not have European-style legislation regarding job security. Although there are some court protections against “wrongful discharge” and discharges related to race, sex and other categories, non-union United States employers are generally free to make employment decisions. However, as noted earlier, there is an inflexibility with regard to nominal wage cuts. These can occur – no law prevents them so long as the minimum wage is paid – but, empirically, worker morale is hurt by such cuts and employers seek to avoid them. Thus, employers in the United States (and other countries with relatively low levels of regulation) face a risk in offering a nominal wage that may be difficult to reduce in the event of falling demand. Under such circumstances, they “insure” themselves against downward wage-rigidity risk by offering a lower wage and paying a cost in the form of increased worker turnover and vacancies. Such a lower-wage policy, particularly when adopted by many employers, would be sustainable and monopsonistic in its effect.

It should be evident that many plausible reasons exist for thinking that monopsony prevails in non-union labour markets. We may think of employers as setting wages in a non-union environment, either by comparison with other employers or by considering both their own individual labour-supply curve and the inherent difficulties of later lowering wages. Thereby, either individual firms would plausibly behave monopsonistically, or else they could somehow collectively do so (e.g. through the relative wage-setting process and each individual firm’s concerns about what will happen if it pays higher wages than its product market competitors). We argue that monopsony and employer ascendancy are the “competitive norm” once the threat of unionization is largely removed, and that labour economics textbooks presenting monopsony as a pedagogical curiosity and demand = supply as the competitive norm need to be rewritten.

Monopsony in time and place

The monopsony model can be helpful in understanding employment and wage dynamics over time, as well. There is a widespread perception, for example, that United States labour markets had much higher rates of employee turnover before the Great Depression than after the Second World War. Commentators
of the post-war era noted a substantial decline in turnover by the 1950s (Ross, 1958; Jacoby, 1983).

One would expect the pre-Depression regime of non-union monopsony and resultant wage repression to exhibit higher quit rates than the post-war regime of high union or union-influenced wages. A popular view during the Great Depression was that wage repression had led to worker underconsumption and caused the economic decline. Indeed, this theme appears in the pre-amble to the Wagner Act of 1935 as a policy justification for promoting unions. The idea was also a factor in the design of the earlier National Industrial Recovery Act of 1933 and the later Fair Labor Standards Act of 1938 (Kaufman, 1996; Kaufman, 1993, p. 61). Our basic argument is that, absent unions, labour markets tend toward monopsony – a belief with significant currency during the Great Depression.

Union decline has not been confined to the United States by any means. In most developed countries, similar declines have occurred, since the 1980s, albeit at different paces and starting from different levels (OECD, 1994, p. 184; ILO, 1997, pp. 239–240). In some countries, however, direct union pay-determination is more detached from actual union representation by various forms of wage “extension” through which bargained wages are applied more broadly (OECD, 1997, pp. 71–72). Conversion to monopsony may be delayed in such countries despite union decline.

Similarly, the predominantly non-union labour markets in Asia may explain why some of those countries were able to operate with very low unemployment and chronic labour shortages until the Asian financial crisis of the late 1990s. As noted earlier, in many Asian countries, independent unions – particularly militant ones – are discouraged by law or informal public policies (Kuruvilla and Erickson, 2002). Officially sponsored unions – where they exist – do not generally push up wages aggressively. They are expected instead to cooperate with management and to keep wage costs from impinging on export-led growth strategies of the authorities.

The Asian financial crisis was sufficiently large to prevent monopsonistic labour markets from completely absorbing the negative employment impact. But monopsony may explain why the Asian financial crisis was comparatively short-lived and why reports of labour shortages resumed after the crisis receded.13 The quick recovery from the financial shock matches our macro predictions.

The NAIRU debate

A mainstay in macroeconomics since the 1960s has been a concept known as the NAIRU, or “non-accelerating inflation rate of unemployment,” sometimes also

12 Unfortunately, continuous United States collection of employee turnover data ended in 1981.
13 A Nexis/Lexis article search for “labour shortage” and each of the “four-tiger” countries (Singapore, Hong Kong, South Korea, Taiwan) produced the following results: 1995: 538; 1999: 526; 1996: 563; 2000, 730; 1997, 608; 2001, 231; 1998, 376. Thus, while there is a dip in labour shortage citations during the Asian financial crisis and the United States recession of 2001, shortages continued to be reported, even during those periods. See Erickson and Kuruvilla (1998a) for a preliminary discussion of industrial relations implications of the crisis.
Monopsony as a metaphor for the emerging post-union labour market

termed the “natural rate of unemployment” the idea that once the unemploy-
ment rate falls below a particular level, inflation will gradually accelerate. A
corollary is that an unemployment rate below the NAIRU is ultimately unsus-
tainable. As late as the mid-1990s, key policy-makers at the Federal Reserve
apparently assumed that the NAIRU stood at about 6 per cent and thus tight-
ened monetary policy when unemployment seemed likely to drop below that
level (Gordon, 1997, p. 12).

However, the performance of the United States economy and labour
market since the late 1990s convinced more recent observers that the NAIRU
has dropped – but also that determining the cause of that decline was difficult
(Staiger, Stock and Watson, 1997; Ball and Mankiw, 2002). Moreover, apart
from the NAIRU, the United States economy seemed to exhibit more stability
of output after the two back-to-back recessions of the early 1980s (Blanchard
and Simon, 2001). This greater stability – whether measured by variations in
unemployment or real GDP – was particularly evident in the 1990s (Mankiw,
2001, pp. 17–18). The recession of 2001 was remarkably mild, particularly in
view of the bursting of a stock market bubble – especially in the technology
sector – the accompanying fall in business investment and the economic reper-

A variety of explanations have been put forward for the decline of the
NAIRU, ranging from change in technology (Card and Dinardo, 2002; Mishel,
Bernstein and Schmitt, 1999, pp. 197–207), to productivity growth (Ball and
Moffitt, 2001; Woodward, 2000, pp. 172–175; Mankiw, 2001, 24), to traumitiza-
tion of workers (Woodward, 2000, pp. 168–169), to demographic and institu-
tional changes (Stiglitz, 1997, pp. 6–7; Western and Beckett, 1999; Autor and
Duggan, 2002; Tulip, 2000; Otoo, 1999; Cohen, Dickens and Posen, 2001). The
union decline and monopsony story is uniquely consistent with the stylized
facts we mentioned above: the reduced use of lay-offs, the widening of earnings
inequality (particularly if monopsony is most likely at the low-wage end of the
labour market), the drop in labour force participation, and the proliferation of
labour shortages, in addition to the apparent decline in the NAIRU.

Moreover, while the NAIRU in the United States may have declined rela-
tive to those of other countries in Europe and elsewhere in the developed
world (Bertola, Blau and Kahn, 2001), owing to factors such as the wage exten-
sion phenomenon, we would expect the same basic trend wherever unions are
losing ground.

Can we have good macro performance
without losing employee voice?

The de-unionization of the United States labour market in particular was
accompanied by new macroeconomic conditions, consistent with monopsony.
But this linkage raises a disturbing implication. The old Phillips curve sug-
gested a trade-off between inflation and unemployment. Do we now replace it
with a new curve that trades off employee voice for low unemployment? And
in the United States context, more than voice is at stake. De-unionization has
likely also played some role in widening wage inequality and in the decline of health insurance and defined-benefit pension coverage, among many other losses for working people.

Recent research has found a latent employee demand among non-union employees in the United States and the United Kingdom for greater voice at the workplace (Freeman and Rogers, 1999; Belfield and Heywood, 2004). In some cases, this latent demand involves traditional unions; in other cases something more like a European-style works council or co-determination that involves participation in firm strategic decision-making. Note, however, that union and worker involvement in firm strategic decision-making would very much go against the tradition of job-control unionism, and would represent a major change (and perhaps even a “transformation”) of the New Deal system of industrial relations (Piore and Sabel, 1984; Kochan, Katz and McKersie, 1986; Erick-son and Kuruvilla, 1998b). In any event, employers are likely to underprovide voice under monopsony, just as they underprovide wages and other benefits.

Conclusion

If one starts with the proposition that perfect competition (demand = supply) is the model for the non-union sector, then it can be argued that whatever emerges in the employment relationship is as much the product of employee desires as employer policies. Workers want to be contingent because of the flexibility that status provides. They prefer defined-contribution plans to defined-benefit pensions because they can invest their own funds. They prefer to “manage their own careers” because they do not want to be dependent on employers. And if the risk of lay-off is the price workers must pay for being their own career managers, so be it. Textbook-perfect competition suggests that this is the best of all possible worlds.

If, however, the employment relationship is the result of monopsony and employer ascendancy, all of these propositions that became popular in the 1990s and after are open to question. The notion of an inherent imbalance of bargaining power revives with regard not only to wages but also to the entire range of employment conditions. Countries experiencing de-unionization (for example, the United States) need to find a better balance between employee voice and macroeconomic performance. The monopsony model of the labour market suggests that, in this respect at least, improvements to the existing world may be possible.

References


Coy, Peter; Ewing, Jack. 2007. “Where are all the workers?”, in *Business Week*, 9 Apr.


—. 1986. “Expansion without inflation: We can have it in the late 1980s”, in UCLA Business Forecasting Project: The UCLA National Business Forecast (June), pp. 35–42.


