Exploitation in Monopsony

Chung-Cheng Lin

Institute of Economics, Academia Sinica, Taipei, Taiwan
Email: cclin@econ.sinica.edu.tw

Received 16 June 2015; accepted 2 August 2015; published 5 August 2015

Abstract

A key feature of monopsony model is that a single firm pays its workers a wage ($w$) less than the marginal revenue product (MRP). This feature has been explained as a synonym of the single firm exploiting its workers since its creation by Joan Robinson [1]. By using a simple standard efficiency wage model of Yellen [2], this paper examines the conventional wisdom by showing that the firm pays workers $w < MRP$ in the equilibrium of full employment, but paradoxically pays them $w = MRP$ in the equilibrium of involuntary unemployment. According to the conventional wisdom that the result of $w < MRP$ implies that workers are exploited by the firm, this finding indicates that the firm does not exploit its employees ($w = MRP$) when there are involuntary unemployed workers queuing for jobs, but paradoxically exploits workers ($w < MRP$) when there are no workers queuing for jobs. The finding is obviously counter-intuitive. This counter-intuitive finding reveals that the key feature of $w < MRP$ in monopsony cannot be regarded as a proper theoretical basis for the issue of labor exploitation.

Keywords

Monopsony, Exploitation, Efficiency Wages

1. Introduction

The origin of the theoretical analysis of monopsony can be traced back to the Joan Robinson [1] classical work—The Economics of Imperfect Competition. Monopsony originally refers to a market structure in which there is only a single buyer. In a survey paper, Boal and Ransom [3] consider that the term “monopsony” nowadays can be applied more broadly to any model where each individual firm faces an upward-sloping labor supply curve and has the market power to set the wage, no matter how many buyers there are in the labor market. Simi-

larly, in a book, Manning [4] extends the scope of monopsony to include models where there are important frictions in the labor market and employers set wages. Manning [4] claims that “the main advantage of the monopsonistic approach is that the way one thinks about labor markets is more ‘natural’ and less forced”. Manning [4] advocates “the belief that a perspective on labor markets based on the view that ‘monopsony’ is important led to a much better understanding of a very wide range of labor market phenomena”.

The main feature of monopsony is that the firm pays its employees a wage (w) that is less than the marginal revenue product (MRP). Joan Robinson [1] interprets this result as a case where the employer exploits employees, and indicates that the reason for the exploitation “is due to imperfection of the labour market” (p. 299). Arthur Pigou [10] uses the value \( \frac{(MRP - w)}{w} \) to measure the degree of exploitation. Many other authors, following Robinson and Pigou, have examined to what extent workers are exploited by a single employer or a few employers. Most of the research interest focuses on whether or not the multi-million-dollars-a-year professional athletes are exploited, and the empirical findings indicate that they are always underpaid. Recently, monopsony has been used to explain a variety of facts that are difficult to explain in the competitive labor model, such as the positive employment effect of minimum wages, the provision of general training, wage dispersion, and employment size-wage effect.

This paper examines the exploitation explanation in a standard efficiency wage model of Yellen [2]. In this simple efficiency wage model, there is a single firm having the market power to set the wage (the same assumptions as in the Robinson monopsony model), and the labor market equilibrium may be characterized by either full employment or involuntary unemployment (the full-employment Robinson model is an extreme case of this model). This study shows that when the labor market equilibrium results in full employment, workers are paid less than the marginal revenue product (w < MRP). The result is the same as that of the (full-employment) Robinson model, the so-called exploitation occurs. By contrast, when the labor market is in the equilibrium of involuntary unemployment, workers are paid according to their marginal revenue product (w = MRP). In this non-full-employment case, even though there is only a single buyer, the so-called exploitation does not occur.

According to the conventional wisdom that the result of w < MRP implies workers are exploited by the firm, we find that the firm does not exploit its employees when there are involuntary unemployed workers queuing for jobs, but paradoxically exploits them when there are no workers queuing for jobs. Since the wage bargaining power of workers is weaker in the case of involuntary employment than in the equilibrium of full employment, the finding is obviously confusing and counter-intuitive. This paradoxical result motivates this paper to reexamine the validity of exploitation explanation and leads to the conclusion that the key feature of w < MRP in monopsony cannot be considered to be a solid theoretical basis for the issue of labor exploitation.

The rest of this paper proceeds as follows. Section 2 presents and analyzes a standard efficiency wage model. Section 3 discusses the validity of exploitation viewpoint in monopsony. A final comment is presented in Section 4.

There are many recent studies involving the labor monopsony model, for examples, Ashenfelter et al. [6], Benson [7], Cahuc and Laroque [8], and Gangopadhyay and Shankar [9].

Robinson [1] defines “exploitation” as “[w] hat is actually meant by exploitation is, usually that the wage is less than the marginal revenue of labour…” . For further discussions, see Chapter 25 of Robinson [1].

Robinson [1] also refers to “the type of exploitation which arises because the supply of labour is imperfectly elastic to the unit of control”.

Alternatively, Robinson [1] points out “exploitation occurs as a result of imperfection in the supply of labour”.

Prior to the 1970s, professional athletes (baseball, basketball, and football) faced significant restrictions on their mobility. The uniform contract contained a reserve clause that virtually bound the player to his team. The reserve clause permitted the team to renew the player’s contract unilaterally year after year, or to transfer him to another team in exchange for cash or another player. Under the reserve system, team owners enjoyed considerable monopsony power; players either had to bargain with the team holding their contracts or leave professional sports. In 1976, a new contract was signed in major league baseball, which made it easier for players to become free agents. Free agency provides players with the opportunity to sell themselves to the highest bidder. This and other later developments have lessened the monopsony power of the team. See Rottenberg [11], and Gilroy and Madden [12] for further discussion. In a recent textbook regarding sports economics, Downward et al. [13] used the monopsony model as the framework to explain why players may be exploited.

Scully [14] and Medoff [15] have indicated that labor market restrictions led to the exploitation of professional baseball players. Raimondo [16], and Sommers and Quinton [17] have found that the removal of these restrictions resulted in salaries reflecting a baseball player’s value to his team (team worth). See Manning [4] for further discussions.

The key difference between this simple standard efficiency wage model and Yellen’s is that a single firm setting is adopted in this paper to simplify the many identical firms assumption in Yellen’s. Supposing that there are many identical firms in this model will make the analysis a little more complex without altering the conclusion. Since the original Robinson monopsony is a single firm model, for the sake of comparison, this paper assumes there is only a single firm.
2. The Model

The basic tenet of the efficiency wage theory is that the effort (productivity) of a typical worker is positively related to his/her wage, and firms have the market power to set the wage. Since a lower wage implies lower productivity, it may be profitable for firms not to lower their wages in the presence of involuntary unemployment. The theory is generally regarded as a plausible explanation as to why involuntary unemployment has persisted in the labor market. The standard or rudimentary efficiency wage model proposed by Yellen is the simplest version of the theory. Its essential feature is that the effort function is directly specified without providing any micro-foundation by examining workers’ optimization behavior. This rudimentary type of efficiency wage model is always used as the analytical framework to illustrate the essential reason why wages do not fall to clear the labor market in the presence of mass unemployment. Some authors have used the standard efficiency wage model to explore various economic issues, such as Schmidt-Sørensen; and the static model of the efficiency wages can be easy to extend to a dynamic model such as Lin and Lai, Faria, Jellal and Zenou.

Let us consider a monopsonistic firm hiring a number of workers to produce a single product. The firm chooses the number of employees \( n \) and the wage \( w \) to maximize its profit \( \pi \), subject to the condition that the quantity of labor demanded \( n \) must not more than the quantity of labor supplied \( N(w) \) in the labor market. That is:

\[
\max_{n,w} \pi = f(e(w)n) - wn; f' > 0, f'' < 0, e_w(w) > 0 \\
\text{s.t.} \quad n \leq N(w); N_w(w) > 0
\] (1)

(2)

The effort function \( e(w) \) with \( e_w(w) > 0 \) in Equation (1) states that the effort of a typical worker is positively related to his/her wage. This is the basic belief of the efficiency wage theory. The production technology \( f(\cdot) \) is a positive function of the effective labor force \( en \) with diminishing marginal returns \( f' > 0 \) and \( f'' < 0 \). The upward-sloping labor supply \( N(w) \) with \( N_w(w) \) in Equation (2) reveals that the higher the wage, the more the workers want to find jobs in the labor market.

2.1. The Robinson Monopsony

For ease of comparison, Robinson’s monopsony can be regarded as an extreme case of this standard efficiency wage model when \( e_w(w) = 0 \) and \( e(w) = 1 \). In this case, the firm’s corresponding first-order condition associated with the wage \( w \) is:

\[
\pi_w = e_w(w)nf'(e(w)n) - n = -n < 0
\] (3)

This result \( \pi_w < 0 \) indicates that, if the workers’ productivity has nothing to do with their wage \( e_w(w) = 0 \), given any level of employment \( n \), a higher wage will always decrease the firm’s profit by raising the labor cost, the firm will thus set the wage as low as possible. It is obvious that the lowest wage that all workers are willing to accept for the job is the marginal workers’ opportunity cost. The marginal workers’ opportunity cost is the reservation wage (i.e. the height of the market labor supply curve) that marginal workers are willing to accept to enter or stay in the labor market. As a result, the firm will set the wage along the market labor supply curve and the labor market equilibrium will result in a state of full employment.

In this case, the monopsonistic firm’s optimization problem is to choose its optimal employment \( n \) to maximize \( \pi = f(n) - wn \) subject to \( n = N_w(w) \). It is equivalent to saying:

\[
\max_{N} \pi = f(N) - Nw(N); w_N(N) > 0
\] (4)

where \( w(N) \) with \( w_N(N) > 0 \) is the inverse function of market labor supply \( N(w) \). The corresponding
first-order condition with respect to the employment $N$ is:

$$
\pi_N = f'(N) - w - Nw_N(N) = 0
$$

Equation (5) can be expressed in another way as:

$$
M\text{RP}_\pi = f'(N) = w + Nw_N = MC
$$

This result is demonstrated in Figure 1. Equation (6) states that a profit-maximizing firm will set its optimal employment $(N^*)$ to the level where the workers’ marginal revenue product $(M\text{RP} = f'(N^*))$ equals the marginal (labor) cost $(MC = w + Nw_N)$. By putting the optimal $N^*$ into the inverse labor supply curve $w(N)$, the firm’s optimal wage offer can be solved as $w^* = w(N^*)$ which is less than $M\text{RP} = f'(N^*)$.

As a result, we have already shown the well-known result in the monopsony model whereby a single buyer will pay its workers a wage $w^* = w(N^*)$ less than their marginal revenue product. It is worth noting that the market equilibrium point $(E)$ of employment and wage $(N^*, w^*)$ is exactly on the labor supply curve. There are $N^*$ workers whose individual reservation wage is not higher than the wage $w^*$, and those who are willing to work at the wage and with all being hired in equilibrium. The labor market equilibrium is therefore characterized by full employment. Accordingly, we obtain:

**Proposition 1:** When the workers’ effort has nothing to do with their wage, then the labor market equilibrium will result in a state of full employment, and the firm will pay its employees a wage less than their marginal revenue product.

Why will the monopsonist pay a wage $w^* = w(N^*)$ that is less than $M\text{RP}$ by the amount $N^*w_N$ when the labor market is in the state of full-employment equilibrium? The amount of underpayment $N^*w_N$ is the product of the employment $(N^*)$ and the slope of the market labor supply curve $(w_N)$. The slope $w_N$ is the increasing rate of the wage as the employment expands. Why, then, does the single labor buyer have to increase its wage offer as it expands its labor force? This is because the single firm pays marginal workers their reservation wage, and so the labor supply curve is binding (labor market exhibits full employment). In order to attract and hire each additional worker with a higher reservation wage “progressively higher wages have to be paid to all in order to attract fresh supplies of labour” (Robinson [1]). Since all the employed workers are paid the same wage, an additional hire will drive up the wage of all infra-marginal workers by the amount of $N^*w_N$. The marginal cost of marginal workers is thus $MC = w^* + N^*w_N$ rather than the wage $w^*$ that they are paid. The optimal decision rule of a profit-maximizing firm is to set its employment at the level where the marginal condition $(M\text{RP} = MC)$ is satisfied. Since $MC = w^* + N^*w_N$ is greater than the wage $w^*$, the firm’s optimal choice, accordingly, occurs in the situation where $w^* < M\text{RP}$ rather than $w^* = M\text{RP}$.

---

12The second-order condition is satisfied since $\pi_{ss} = f''(N) - 2w_N - Nw_{ss} < 0$. 

---

Figure 1. Robinson’s monopsony.
2.2. The Standard Efficiency Wage Model

When \( e_w(w) > 0 \), the firm’s goal is to maximize the profit \( \pi \) in Equation (1) by choosing its employment and wage. By ignoring the constraint in Equation (2) for the time being, the first-order conditions associated with \( n \) and \( w \), respectively, are:

\[
\pi_n = e(w) f'(e(w)n) - w = 0 \tag{7}
\]

\[
\pi_w = e_w(w) nf'(e(w)n) - n = 0 \tag{8}
\]

Equation (7) states that, given the wage, the quantity of labor is employed at the level where the marginal revenue product of labor \( ef' \) equals the marginal cost of labor \( w \). Equation (8) indicates that, given the quantity of labor, the wage is set at the level where the marginal revenue from the wage \( e_w nf' \) equals the marginal cost of the wage \( n \). The second-order conditions require that:

\[
\pi_{nn} = e^2 f'' < 0 \quad \text{and} \quad \pi_{wn} \pi_{ww} - \pi_{wn} \pi_{ww} = e^2 e_{ww} nf'' f'' > 0 \tag{9}
\]

The firm’s notional number of workers employed \( n^n \) and the notional wage offered \( w^n \) are determined by solving Equations (7) and Equation (8)\(^{13} \).

The \( NN \) curve and the \( WW \) curve represent the optimal choice of the firm’s labor demand and wage-setting conditions in Equation (7) and Equation (8), respectively. Their slopes are both negative because:

\[
\frac{dw}{dn}_{NN} = -\frac{\pi_{wn}}{\pi_{ww}} = -\frac{e^3 f^n}{ee' nf^n} = -\frac{e}{e'n} < 0 \tag{10}
\]

\[
\frac{dw}{dn}_{WW} = -\frac{\pi_{wn}}{\pi_{ww}} = -\frac{e e' nf^n}{e' nf^n + (e')^2 n^2 f^n} < 0 \tag{11}
\]

Moreover, the \( NN \) curve is steeper than the \( WW \) curve due to the second-order conditions of the firm’s optimization in Equation (9) where \( \pi_{nn} \pi_{ww} - \pi_{ww} \pi_{wn} = e^2 e_{ww} nf'' f'' > 0 \). The firm’s notional number of workers employed \( n^n \) and the notional wage offered \( w^n \) are determined by the intersection of both curves.

Let the positively-sloping line \( \overline{NN} \) represent the upward-sloping labor supply \( N(w) \) with \( N_w(w) > 0 \) in Equation (2). By drawing all of three \( \overline{NN} \), \( NN \) and \( WW \) curves together in a figure, we can discuss whether or not the firm’s notional levels \( (n^n \text{ and } w^n) \) will turn out to be the realized equilibrium levels \( (n^* \text{ and } w^*) \) or not.

2.2.1. The Market Labor Supply Curve Is Not Binding

In Figure 2, the point of intersection \( (E) \) of \( NN \) and \( WW \) is located to the left part of \( \overline{NN} \). There are

---

\(^{13}\)The term “notional level” here refers to the firm’s optimal levels of labor demanded and wage offer without taking into account the resource constraint in Equation (2). When the constraint is not binding, the notional level will be the realized level, too; otherwise, they will not the same. For further discussions, see the following analysis.
workers whose reservation wage is not higher than the wage \( w^* \). They are willing to find a job at the wage, but the total available jobs the firm intends to afford is only \( n^* \). We have \( N\left(w^*\right) > n^* \). There are \( N\left(w^*\right) - n^* \) workers queuing for jobs, but they are not hired and thus become unemployed involuntarily. In this case of an excess supply of labor, the firm will not face the scarcity constraint of manpower at the notional wage \( w^* \). The notional levels of \( n^* \) and \( w^* \) will be the realized equilibrium levels as well.

A question that naturally arises is why the firm does not cut its wage in the presence of unemployed workers who are willing to work as long as the market wage is higher than their reservation wage. Alternatively, why does the firm not cut the wage when the wage offer is higher than the marginal workers’ opportunity cost? In fact, the firm does intend to cut its wage if it is profitable. This becomes clear when we consider the conventional setting in monopsony where \( e_\cdot(w) = 0 \) and \( e'(w) = 1 \). In this case, Equation (8) becomes Equation (3) where \( \pi_w = -n < 0 \), and the wage will be set as low as possible until the wage equals the marginal workers’ opportunity cost. However, when a higher wage implies higher productivity \( e_\cdot(w) > 0 \), there is a benefit as well as a cost to a firm associated with paying a higher wage. The firm may find it is profitable not to cut its wage to eliminate the excess supply of labor.

From Equation (7), it is very easy to see that the optimal choice of the monopsonistic firm is to pay its workers a wage \( MC = w \) that is just equal to their marginal revenue product \( MRP = e' \). We thus conclude:

**Proposition 2:** When workers’ effort is positively related to their wage, and if the labor market equilibrium is characterized by a state of involuntary employment, then the firm will pay its employees a wage \( w \) equals their marginal revenue product \( MRP \).

### 2.2.2. The Market Labor Supply Curve Is Binding

In Figure 3, the point \( E \) at which the \( NN \) and the \( WW \) curves intersect is located to the right side of \( NN \). The total number of available jobs that the firm wants to afford at the wage \( w^* \) is \( n^* \). However, there are only \( N\left(w^*\right) \) workers whose individual reservation wage is not higher than the wage \( w^* \) and they are willing to find a job at that wage. We thus have \( N\left(w^*\right) < n^* \). There are \( n^* - N\left(w^*\right) \) job vacancies that the firm wants to afford at the wage \( w^* \) but it cannot recruit more workers to fill these vacancies. In this case where there is an excess demand of labor (a labor shortage), the firm will face the scarcity constraint of manpower at the notional wage \( w^* \). The notional levels of \( n^* \) and \( w^* \) will not be the realized levels.

In this labor shortage situation, the firm must take the constraint \( n = N\left(w\right) \) or \( N = N\left(w\right) \) into account when determining the employment and the wage levels that can be realized in practice. The firm that maximizes its profit \( \pi = f\left(e\left(w\left(N\right)\right)\right) - Nw\left(N\right) \) will choose the optimal employment \( \left(N^*\right) \) at the level where

![Figure 3. Excess demand of labor.](image)

\( ^{14} \)This is a standard result in the efficiency wage models. For example, Yellen [2] and Akerlof and Yellen [25] have pointed out that: “Each firm should then optimally hire labor up to the point where its marginal product, \( e\left(w^*\right) f'(e\left(w^*\right) n^*) \), is equal to the real wage \( w^* \).”
Since the labor supply curve is binding, the labor market equilibrium will result in a state of full employment. In this case where \( MRP = ef' + e_Nf'w_N \) and \( MC = w + Nw_N \), the well-known result \( w < MRP \) appears as in Robinson’s monopsony. The underlying economic intuition is the same as that for the full-employment Robinson monopsony.

**Proposition 3:** When workers’ effort is positively related to their wage, and if the labor market equilibrium is characterized by a state of full employment, then the firm will pay its employees a wage that is less than their marginal revenue product.

Summarizing the above three propositions, we conclude that regardless of whether workers’ productivity is positively related to their wage or not, the firm pays them \( w = MRP \) when there are involuntary unemployed workers queuing for jobs, but paradoxically pays workers \( w < MRP \) when there is full employment.

In sum, according to the conventional wisdom that the result of \( w < MRP \) implies workers are exploited by the firm, then we find that the firm does not exploit workers when there are workers queuing for jobs (involuntary unemployment), but exploits them when there are no workers queuing for jobs (full employment or a labor shortage). Since the wage bargaining strength of workers is stronger in the case of full employment than the equilibrium of involuntary employment, this finding is obviously counter-intuitive. It implies that the conventional wisdom regarding exploitation in monopsony is not convincing.

### 3. Conclusions

By using a simple standard efficiency wage model of Yellen [2], this paper explores this conventional wisdom. Just as in the case of Robinson’s monopsony, the setting of this standard efficiency wage model is that there is a single firm and the firm has the market power to set the wage. Differences from Robinson’s monopsony in which the labor market equilibrium must be in full employment, the equilibrium may be characterized by either full employment or involuntary unemployment in this standard efficiency wage model. This paper finds that the single firm pays \( w = MRP \) when there are workers queuing for jobs (involuntary unemployment), but pays \( w < MRP \) when there are no workers queuing for jobs (full employment or a labor shortage). This finding is obviously counter-intuitive and motivates the author of this paper to obtain a deeper understanding of the real reason for \( w < MRP \) in monopsony.

At first glance, this result is confusing and presents us with a conundrum. Further inspection indicates that the reason for \( w < MRP \) is not merely due to the unique market structure of monopsony, but also because the firm faces a constraint in terms of the scarcity of manpower (a binding labor supply curve). Since all the employed workers are assumed to be paid the same wage, when the market labor supply curve is binding (i.e. there is full-employment), the firm must offer a successively higher wage to attract and employ each additional worker along the upward-sloping labor supply curve. An extra or a new hire will generate a positive “externality” with respect to all of the infra-marginal workers by driving up the wage offered. This positive externality enjoyed by infra-marginal workers is an extra cost or burden to the firm. In taking this into account, the single firm chooses an employment-wage package such that the difference between the marginal workers’ marginal revenue product and the wage is just enough to cover the extra cost. In other words, the optimal choice of package occurs where the wage equals the marginal workers’ net or real contribution to \( MRP \) (i.e. the \( MRP \) minus the extra cost).

If the labor supply curve is not binding (i.e. there is an involuntary unemployment), the firm can hire an additional worker without incurring the extra cost. The firm does not need to generate a gap between the marginal workers’ marginal revenue product and the wage to pay the extra cost. Without the extra burden, any gap between the wage and the marginal revenue product implies a potential profit that is not yet arbitraged; the firm will expand employment until \( w = MRP \) to let the gap vanish. As a conclusion, whether the single firm will pay workers \( w < MRP \) or \( w = MRP \) is crucially dependent upon the upward-sloping labor supply curve, whether it is binding (the labor market is characterized by full employment) or not.

The finding, that the firm exploits workers when there is full employment but does not exploit them when there are involuntary unemployed workers, is obviously paradoxical. We therefore conclude that the key feature of \( w < MRP \) in monopsony is not a solid theoretical basis for the issue of labor exploitation. A new theory needs to be developed in the future studies.

### 4. A Final Comment

Finally, I agree with Manning [4]’s perspective that “[a]ssuming labor markets are monopsonistic also brings
the thinking of labor economists in line with the way in which agents perceive the workings of labor markets. I also believe Manning’s argument “that our understanding of labor markets is markedly improved by explicit recognition of the fact that employers have some market power in the determination of wages”. In fact, an increasing number of recent studies have been devoted to exploring the potential implications of monopsony. Those who are familiar with the recent monopsony literature may expect that the importance of this paper’s finding (i.e. a monopsonistic firm may pay workers $w = MRP$, or the upward-sloping labor supply curve may be not binding) may go beyond the issue of exploitation. For example, some studies consider that the greater the value $(MRP - w)/w$, the greater monopsony power. A number of other studies claim that a binding positively-sloping labor supply curve in monopsony may explain why the minimum wage legislation may increase employment, and that a firm hiring more workers always pays higher wages as well (the employment size-wage effect). When we recognize that a monopsonistic firm may pay workers $w = MRP$ (or the upward-sloping labor supply curve may be not binding), perhaps more attentions needs to be paid when exploring these issues.

References


15Some relevant studies are listed as follows: Hyde and Perloff [33], Boal [34], Hirsch and Schumacher [35] [36], Azzam [37], Green, Machin and Manning [38], Boal and Ransom [3], Pauly [39], Stagner, Spetz and Phibbs [40], Bluskar and To [41], Baskar, Manning and To [5], and Manning [4][42][44].
16See Manning [42] for a brief list of relevant applications.


