The Quarterly Journal of Austrian Economics

VOL. 15 | NO. 3 | 277–303
FALL 2012

Efficiency Wages and Involuntary Unemployment Revisited

Pavel Ryska and Jan Průša

ABSTRACT: In this paper we tackle two shortcomings of the present efficiency wage models. Firstly, they do not fully account for labor heterogeneity, thus implying that high-effort and low-effort units of labor are interchangeable. Secondly, building on this assumed homogeneity of labor, the models derive involuntary unemployment from effort decisions of workers, which are patently voluntary.

We offer a consistent reformulation of the theory: Each of the effort or quality levels is regarded as a separate market which has its own clearing quantity and price. As such unemployment is a result of workers' reluctance to adjust to the prevailing market conditions on the respective labor sub-market.

To further clarify heterogeneity in labor markets, we propose to employ the demand for workers' characteristics instead of the demand for workers. This microeconomic approach shows that in standard equilibrium, employers will not choose among all workers but only select specific characteristic-types. Therefore, to become attractive, an unemployed worker has to significantly alter either his wage or the bundle of offered characteristics.

Pavel Ryska (pavel_ryska@volny.cz) and Jan Průša (jan.prusa@ies-prague.org) are both of the Institute of Economic Studies, Charles University in Prague, Czech Republic. The support from the Grant Agency of Charles University (GAUK) under projects 684012, 89910 and SVV 265504, and from the Grant Agency of the Czech Republic (GACR) under projects P402/11/0948 and P402/12/0982 is gratefully acknowledged. The views expressed here are those of the authors and not necessarily those of our institution. All remaining errors are solely our responsibility.

Both of these modifications reinforce our central claim that free market interaction cannot lead to unemployment other than voluntary.

KEYWORDS: labor market, efficiency wages, involuntary unemployment, demand for heterogeneous goods

JEL CLASSIFICATION: J64, J20

1 INTRODUCTION

To understand why there is unemployment and whether it may be involuntary is an important and long-lasting goal of the economic profession. Since the 1980s, the efficiency wage theories have occupied a prominent place in labor market research. Their appeal rests in two features. First, efficiency wage considerations derive the occurrence of involuntary unemployment only from simple profit maximization of the firm. Second, they explain why wages are observed to vary relatively little and employment to vary a lot.

In the present paper, we take issue with the main conclusions of efficiency wage models. In our view, these models incorrectly specify the labor market as a homogeneous one, while they implicitly model a situation of heterogeneous labor. This is why they necessarily lead to a non-Walrasian outcome and imply involuntary unemployment. We show that if labor markets are modeled as fully heterogeneous—an approach that we support—then the existence of involuntary unemployment does not follow. Specifically, we analyze the two most common efficiency wage models: the Solow (1979) model, in some literature called the "generic" efficiency wage model, and the Shapiro-Stiglitz (1984) "shirking" model. Although the two models differ in the degree of wage rigidity they lead to, they both imply involuntary unemployment and are both based on the same microeconomic approach to labor markets. We offer an alternative approach, and for this purpose we invoke the rather little known contribution of Lancaster (1966, 1971) to show that the conclusion about involuntary unemployment does not hold.

It is important to stress that while our paper does not dispute the existence of unemployment in general, we object to the claim that *involuntary* unemployment is caused by market optimization of firms. Therefore the main purpose of our paper is to contribute to logically consistent theory and model building.

This paper is structured as follows. In Section 2, we describe the two major efficiency models (Solow and Shapiro-Stiglitz) and for each provide critical analysis of its conclusions. In this part, we focus on the particular layout of the model at hand. In Section 3, we generalize our findings and show their deeper implications (subsection 3.1), briefly discuss the limited usefulness of the hedonic wages approach (subsection 3.2), and then provide a formal microeconomic setting (Lancaster's approach) which is adequate to address the problems and which reinforces our case (subsection 3.3). Section 4 concludes the paper. In the appendix, we show how our approach refutes the popular claim by Raff and Summers (1987) that Henry Ford's factory caused involuntary unemployment in the early twentieth century by suddenly doubling the wages it paid.

2 TWO EFFICIENCY WAGE MODELS

2.1 Solow's Model

2.1.1 The Standard Layout

Solow (1979) was one of the first to formulate an efficiency wage model that would imply wage rigidity and therefore involuntary unemployment. Here we use the most common version of the model as presented in Romer (2006) to maintain the familiar notation.

There is a large number of identical competitive firms that maximize profits. To simplify the analysis, only one input—labor—is assumed. The contribution of labor to output of the firm depends positively on effort e that workers exert. This effort is assumed to enter the production function multiplicatively with labor. The production function is therefore

$$(1) \quad Y = F(eL)$$

where $F'(\bullet) > 0$, $F''(\bullet) < 0$. The key element of the model is the assumption that higher wages induce higher effort on worker's part, i.e. that $e=e(w)$, where $e'(\bullet)>0$. Since firms are free to choose both the wage and amount of employment, their optimization problem is

$$(2)^1 \quad \max_{L,w} \{F(e(w)L) - wL\}.$$

[1] We assume that output price is normalized to one, i.e., output is the numeraire.

The first-order conditions are

$$F'(e(w)L)e(w)-w=0$$

and

$$F'(e(w)L)Le'(w)-L=0.$$

Assuming that second-order conditions are satisfied, the two equations yield the solution to optimal choices of the firm. Their combination gives

$$(3) \quad e'(w^*)=\frac{e(w^*)}{w^*}.$$

The interpretation is straightforward. The firm wants to find such w^* so as to minimize the cost per unit of effective labor $w/e(w)$ or, equivalently, maximize effort per dollar $e(w)/w$. This happens at w^* where average effort equals marginal effort which is the condition in (3).[2] This equation therefore gives the optimal wage w^*, which can then be plugged into either of the two first-order conditions to find optimal employment L^*.

The essential message of equation (3) is that optimal wage w^* is completely independent of the labor market. What drives wages is not standard supply of labor and demand for labor, but only the effort function $e(w)$ that workers have and that firms know about. Therefore, the implication of the Solow model is that if the entry of effort into the production function is labor-augmenting (i.e. e multiplies L rather than other potential inputs), then the optimal wage can be completely rigid. This has important consequences. First, because all firms are identical, each chooses w^*, which implies a potential disequilibrium on the labor market. Whenever

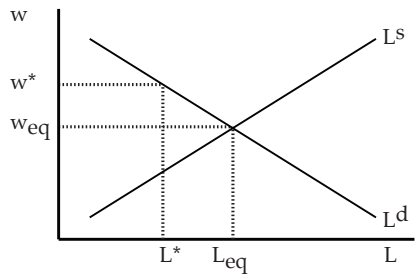
[2] Minimizing real wage costs can be formally written as

$$w^*=\arg\min_w\left\{\frac{w}{e(w)}\right\}.$$

Differentiating with respect to w and assuming that regularity conditions are satisfied, this yields the same solution as in (3).

w^* exceeds the Walrasian equilibrium wage w_{eq}, involuntary unemployment results. This is depicted in Figure 1. Second, because the wage is given by (3) and thus is unresponsive to labor demand, swings in the labor demand curve only affect the amount of labor hired, not the wage level. Therefore, the model seems to explain the empirically observed large movements in employment as compared to the relatively small movements in wages.

Figure 1. Optimal wage w^* in the Solow model



2.1.2 A Closer Look at Assumptions

The above Solow model is striking in that it does not produce disequilibria due to obstacles to market functioning, due to market imperfections or non-maximizing behavior, but precisely *because of* profit-maximizing behavior. The very goal to maximize profits leads firms to disregard the market when setting wages. Solow's model has received repeated references (Yellen, 1984) and extensions (Summers, 1988), and also relatively few criticisms (e.g. Bellante, 1994). However, to the best of our knowledge, the most problematic (and hidden) assumptions in the model have not been discussed.

First, Solow (1979) assumes that the effort function is a fixed, given relationship that cannot be changed by the worker. However, this is not consistent with any conceivable idea of the effort function, nor with definitions of effort by other authors. Second, Solow introduces only one labor market, although he assumes

from the very beginning that there are different types (efforts or qualities) of labor. We discuss these two points in the next two parts, respectively.

2.1.3 The Nature of the Effort Function

The effort function represents the assumption that effort of workers varies with wage, but non-linearly. It is necessary to clarify at the outset that both wages and labor output are flow variables measured per unit of time. Therefore hourly effort $e(\bullet)$ (leading to certain hourly output) is a non-linear function of hourly wage w.

It seems reasonable to assume that marginal effort $e'(\bullet)$ will be decreasing from some point. No matter how high the offered hourly wage, there is an upper limit on how much a human can perform. The assertion of increasing marginal effort at lower wages is more disputed, but can also be justified on grounds of (mal)nutrition in developing countries or motivation in developed countries.

As was shown above, the firm then sets the optimal wage w^* to satisfy the unit wage elasticity of effort in (3). Following this condition, efficiency wage models proceed to assert that this generates downward wage rigidity. Workers cannot offer lower wages in order to find a job: In the region where marginal effort is increasing, lower wage $w < w^*$ would violate the optimum condition of cost minimization for the firm and thus decrease its profitability.

This however contradicts the nature of the effort function. Ever since Shapiro and Stiglitz introduced in their seminal paper "the effort decision of a worker" (1984, p. 435), effort $e(\bullet)$ is considered to be endogenous to the worker.[3] If the worker is not successful at finding a job paying certain desired wage per unit of effort, he receives a clear market signal that his services are too expensive. The worker then has to offer a lower price—that is, he has to turn his effort function upwards around the point of origin. Such a matching process is not only possible, but in our opinion also often observable.

To avoid misunderstanding, we are far from claiming that the concept of effort function is flawed or unusable. On the contrary,

[3] The case of malnutrition is excluded from our consideration as it does not concern developed economies.

it provides at least two important insights into the functioning of the labor market. First, the effort function embodies the notion that labor as a productive input has more dimensions than man-hours. Second, it implies that optimization of firms and workers is accordingly multidimensional and that a firm can employ more of low-intensity man-hours or less of high-intensity man-hours.[4]

Nevertheless, what we do question is the conclusions about market disequilibrium drawn from effort functions. The most important consequence of condition (3) is that firms cannot be expected to set wage primarily in terms of hours. Instead they will optimize production in such a way that unit wage (in money per effort) will be set equal to the marginal product of effort, and the final hourly payment will then be determined by the amount of effort exerted. The firm will plan the amount of effort units it requires. It will then hire the respective amount of labor which will supply this amount of effort, that is, either a lot of cheap low-effort workers or few costly high-effort workers. Let us recall that the price per unit of effort (i.e. "effort wage") in which workers can compete is a voluntary decision of workers based on their preferences. At this point we can already conclude: Given that the effort function is fully determined by workers' preferences, there exists no room for involuntary unemployment.

So far we stressed the voluntary, preference-driven character of the effort function. In the following section we will turn to another crucial implication of the efficiency wage models: the multidimensionality of labor.

2.1.4 Heterogeneous Labor Markets

The production function in (1) has only one input—labor. Usually, the simplification to model only one labor market with labor L and one capital market with capital K is justified. This is because we often assume from the very beginning that there is only one type of labor and one type of capital. Then, the maximization of profit in the standard textbook form $\Pi=F(K,L)-r_K K-wL$ does not pose any problems for analysis.

[4] As a result, it shows that the solution to the profit maximization problem may have more than one solution.

We suggest, however, that the situation with efficiency wage models is fundamentally different. In these models, one deals *by definition* with different types of labor. Specifically, in the above model Solow introduces the possibility of different effort levels e, which implies labor of varying quality. Yet, Solow still models the labor market as a single one with homogenous labor L: See the right-hand term wL in (2). This is, in our opinion, a serious contradiction that prevents standard economic analysis—and standard results accordingly. In what follows, we show that this assumption of a homogenenous labor market, which has gone unnoticed in efficiency literature so far, is the necessary element that allows Solow to derive the existence of involuntary unemployment.

In contrast to the production function (1), let us now define that each level of effort e constitutes a different type of labor. That is, we now distinguish between labors of different qualities according to the effort that the worker puts into it. The more effort is put into work, the higher quality of labor follows. For ease of notation, let us also assume that the number of types (qualities) of labor is finite, i.e. ranging from some inferior labor L_1 to top-quality labor L_N. Solow's production function $Y=F(eL)$ now becomes

$$(4) \quad Y=F(L_1,\ldots,L_N).$$

At first sight, effort e is absent in the production function in (4). This is because its meaning is now subsumed under different *types of labor* $L_1,\ldots,L_N$. L_i corresponds to effort level e_i, which in turn corresponds to wage w_i. It is crucial to note that index i does *not* correspond to workers: more workers can be of the same type L_i and one worker can alter his type from L_i to L_j, as we show in section 2.1.3. (We return to this point in the next section.) Therefore, we have preserved the link between wages and effort—the main idea of efficiency wage models—but we model different efforts in the form of different types of labor. Putting these considerations together, the firm now faces the problem

$$(5) \quad \max_{L_1,\ldots,L_N} \{F(L_1,\ldots,L_N)-w_1L_1-\ldots-w_NL_N\}.$$

The optimization problem in this form clarifies why the treatment of labor matters crucially. Now, the firm still chooses effort

(efficiency) by choosing wage, but formally does it by choosing a specific labor market L_i. The pairs $(w_1,L_1),...,(w_N,L_N)$ represent the particular labor markets, assorted by the quality of labor that is traded on them. Importantly, w_i is the *Walrasian equilibrium wage* prevailing on market for labor of type L_i. As a result, in the optimization problem (5) the firm does not explicitly choose w as in (2), but does it implicitly by choosing labor of quality L_i with its corresponding price w_i.

2.1.5 Equilibrium on the Heterogeneous Market

The implications of this reformulation are far-reaching. Firms still maximize profits by choosing the optimal effort level, but since we now view labor markets as heterogenous, firms do so by choosing the optimal labor market where to hire labor. Finding and setting the optimal wage therefore amounts to choosing a labor market and paying its Walrasian *equilibrium* wage. This means, however, that no disequilibrium wage can occur. Any wage chosen is an equilibrium wage on *some* market—and this market is chosen. The effort function $e(w)$ can then be understood as a set of N wage-quality pairs traded on N markets.

By defining labor as heterogeneous, we have reached a different perspective on the problem at hand: firms choose between types (qualities) of labor $L_1,...,L_N$ just as they may choose between types (qualities) of capital $K_1,...,K_N$. Our reformulation, in contrast to Solow's treatment, is therefore a general one. We have shown that different efficiencies (efforts) can be simply modeled as different types of labor, each of which has its own market. This is a parallel to different types of capital having their own market. Solow, by contrast, assumes that labor of different efficiencies is traded on a single market at some *uniform* price. This is why he inevitably derives a non-standard situation in which the wage is independent of demand and supply, and hence where involuntary unemployment exists. By contrast, defining labor properly as heterogenous shows that there is no room for disequilibrium situations and involuntary unemployment.[5] This refutes the idea of

[5] Equilibrium must be attained insofar as neoclassical assumptions on the individual submarkets i are met. This is not to say that other frictions on these markets do

the Solow model that labor is a peculiar, special input that causes standard market equilibrium to fail.

The confusing feature of labor is that one and the same physical person may exert different efficiencies and therefore be present on multiple labor markets. In economics, however, markets are abstract concepts, not physical places. Thus it does not pose any problem to model different efforts of the same person as different markets. We therefore see no reason for abandoning standard economic analysis, contrary to what is done in the Solow (1979) model.

To clarify that what matters more is services offered by inputs rather than their physical definition, in section 3 we introduce an alternative microeconomic approach, using Lancaster's characteristic-space model.

2.2 Shapiro-Stiglitz Model

2.2.1 The No-Shirking Condition

In a second very known version of efficiency wage models, Shapiro and Stiglitz (1984) elaborated on early literature to include imperfect information. Their basic idea was that when firms cannot perfectly monitor workers' effort, they may choose to pay higher wages to discourage workers from shirking. The point of Shapiro-Stiglitz model is not that higher wages simply induce higher effort, which sufficed as reason in the basic Solow (1979) model. Here, non-Walrasian wages play the role of creating involuntary unemployment, which in turn works as a "worker discipline device": if there was no involuntary unemployment, argue Shapiro and Stiglitz, then shirking would entail no costs, because after being caught and fired, workers would immediately find new jobs. Only involuntary unemployment can "discipline" workers by incurring costs of losing jobs. Therefore, Shapiro and Stiglitz understand non-Walrasian wages as an economy-wide equilibrium solution to the possibility of shirking.

The model assumes that there are only two effort levels: working and shirking. As a result, one can inspect three value functions V_E^N,

not exist. We merely show that effort or quality variations of labor *per se* do not generate disequilibrium.

V_E^S and V_U that stand for the present value of (a) being employed and working, (b) being employed and shirking, and (c) being unemployed, respectively. Defining these and then imposing the condition under which workers do not shirk ($V_E^N \geq V_E^S$) yields the well known no-shirking condition

$$(6) \quad w \geq e + \bar{w} + \frac{e}{q}\left(\frac{b}{u} + r\right) \equiv \hat{w},$$

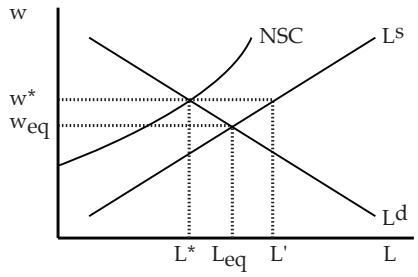
where e is effort (disutility from working), $\bar{w}$ the unemployment benefit (i.e. alternative wage in case of getting fired), q the probability of being caught shirking, b the probability of termination of working post (exogenous, with no relation to effort), u the unemployment rate, and r the rate of interest used for discounting future to present value.

Wage $\hat{w}$ makes workers just indifferent between working and shirking. Any wage higher than $\hat{w}$ ensures that given the current unemployment rate and the parameters of the model, costs of shirking outweigh its benefits. The appealing feature of the model is that the minimum no-shirking wage $\hat{w}$ depends on the unemployment rate in the economy—that is, on other firms' wage decisions—and therefore can be plotted as a function in the labor market diagram. Higher employment L (equivalent to lower unemployment rate u) lowers workers' costs of getting fired for shirking. Therefore, according to (6), to discourage them from shirking, firms have to pay them higher wages. This results in the upward-sloping no-shirking condition (NSC) in Figure 2.[6] The optimal wage that the firm sets is w^*, given by the intersection of the labor-demand curve and the no-shirking condition. This time, unlike the Solow model, labor demand matters for the optimal wage. However, labor supply is still irrelevant in setting of w^*, giving again rise to involuntary unemployment.[7]

[6] Figure 2 is not exactly the one shown in Shapiro, Stiglitz (1984). The authors use a figure with a vertical labor supply curve at the maximum employment N. We prefer to use a more general version with an upward-sloping labor supply curve which nevertheless fully maintains the idea of the Shapiro-Stiglitz model.

[7] More accurately, in the Shapiro-Stiglitz model involuntary unemployment results *necessarily*, because it is precisely the involuntary nature of unemployment that disciplines workers and keeps them from shirking. In the Solow model, by contrast, involuntary unemployment is a possibility, but not a necessity.

Figure 2. Optimal wage *w in the Shapiro-Stiglitz model**



2.2.2 A Play with Words to Prove Involuntary Unemployment

This conclusion about involuntary unemployment was criticized by Carmichael (1985). He suggested that those workers who want to work at *w** but are not hired can simply "buy" the job. That is, they offer to pay an "entrance fee" or a "bond" to the firm for giving them the job. This reduces their value from the job to the point where they are indifferent between working and being unemployed. Therefore, unemployment is no longer involuntary. Although Carmichael's point made an impact,[8] there is a more serious problem in the model.

Shapiro and Stiglitz give the following reasoning: "From the worker's point of view, unemployment is involuntary: those without jobs would be happy to work at w* or lower, but cannot make a credible promise not to shirk at such wages."[9]

In our view, Shapiro and Stiglitz are involved in an outright contradiction. On one hand, they claim that the workers would be "happy to work at *w** or lower." On the other, they assert that at such wages "they cannot make a credible promise not to shirk." One would like to ask the authors the question: "At *w**, would the employees work or shirk?" It appears as if Shapiro and Stiglitz

[8] See Shapiro's and Stiglitz's (1985) reaction in which they retreated somewhat from their claim about involuntary unemployment.

[9] Shapiro and Stiglitz (1984), p. 438, emphasis original.

used a terminological trick on the reader to make their case. The situation in the model is as follows: One can be either employed or unemployed; further, if one is employed, he can work or shirk. If one shirks, he does not work at all and thus does not exert any effort. However trivial this may sound, it matters crucially for the conclusions. Shapiro and Stiglitz have us believe that at w^*, $L'-L^*$ people in fact offer to do something for the potential employer, but somehow cannot get the job. However, as the authors themselves write, they do not offer to work—they instead *offer to shirk*. Therefore, there cannot be involuntary unemployment. When Shapiro and Stiglitz call unemployment involuntary, they in fact pity workers for not getting paid for *not* working.[10]

The confusion of words is caused by the use of the category "employed, but shirking." For economic analysis, it does not matter whether one is idly sitting at home and is called "unemployed," or whether he is idly sitting at work (shirking) and is called "employed, but shirking."[11] The economic meaning of the two categories—the employed but shirking and the unemployed—is identical in the model: the person does not produce anything. Yet, in Shapiro's and Stiglitz's logic, when one offers to shirk at w^* and firms, knowing this, do not offer him the job, he is called "involuntarily unemployed." If Shapiro and Stiglitz had not used the term "employed" for those who do not work at their job, they could never have reached this conclusion.

2.2.3 Labor Supply versus the No-Shirking Condition

Shapiro and Stiglitz chose to introduce the simplest case where the only two options are $e > 0$ for working and $e = 0$ for shirking. Let us broaden the model to a more general case with $0<e_1<e_2$, where

[10] Similar self-contradictory statements like that of Shapiro and Stiglitz above can be found in Yellen (1984, p. 202) and Romer (2006, p. 455).

[11] There is no difference between "unemployed" and "employed, but shirking" in the model since Shapiro and Stiglitz allow only two effort levels in their model: $e > 0$ for working and 0 for shirking. Thus, shirking at work amounts to zero output just as being unemployed does.

As a side note, we add that "employed" in economics has always meant "utilized." Therefore, "employed, but shirking" is a confused connection that the authors fail to explain. See terms such as "employed shirker" in Shapiro, Stiglitz (1984, p. 436.)

e_1 is the shirking state[12] and e_2 the working state. It is easy to show, however, that this does not change anything about the failure of the model to prove involuntary unemployment. It suffices to inspect carefully the definitions of the NSC and the labor supply curve.

In order to construct the no-shirking condition, Shapiro and Stiglitz first had to empty the meaning of the labor supply curve. The labor supply curve, in all economic works since the curve was conceived, has always *meant precisely the no-shirking condition*. The condition of working at a given effort level—as the opposite of shirking (i.e. failing to reach this effort level)—is exactly the idea of the labor supply curve: one is offering to supply labor of given quality in exchange for reward on the labor supply curve, and not to supply this quality beyond the supply curve. Therefore, Shapiro and Stiglitz necessarily have to redefine the labor supply curve away from its original meaning in order to move its content into their newly conceived "no-shirking condition." Hence, in their model, the labor supply curve also comprises those who in fact do not want to supply the labor of given quality. Only then can the authors come up with the NSC curve to denote those who will "really" work. We stress that the invalidity of the model does not rest on whether effort levels are defined as $e_1=0$ and $e_2>0$ or $0<e_1<e_2$. The reason is that the borderline between working and shirking is given by the definition of what L means on the respective market.

At a closer look, the root of the problem just described is the same as that of the Solow model identified in Section 2.1.4: labor L on the horizontal axis in Figure 2 is not always measured in the same units. Labor used in the non-shirking condition is something completely different from "labor" (i.e., not labor but shirking in any form) ascribed by Shapiro and Stiglitz to the labor supply curve. Then the horizontal axis has two alternating meanings, and hence the figure cannot be used for standard microeconomic analysis. We explain deeper problems of this practice in Section 3.

For now, we conclude that there is no involuntary unemployment in the Shapiro-Stiglitz model. Workers voluntarily choose their effort level at any wage. If firms know that with a certain wage

[12] Shirking from the point of view of the firm is any positive effort lower than what was agreed in the employment contract, i.e. any positive effort lower than the marginal product equivalent to the agreed wage.

and certain level of employment additional workers would shirk at work, and for this reason do not offer them work, then nothing of involuntary nature occurs. The NSC curve in reality *coincides with* the correct labor supply curve.[13]

2.2.4 Firm Optimization

The central claim of the Shapiro-Stiglitz model is that firm optimization leads to involuntary unemployment, and that this in turn represents market inefficiency which needs to be corrected. Still the model leaves the interaction of firms on the market largely undefined. Let us therefore analyze the production side in more detail.

The model assumes that there are M identical firms which have aggregate production function (see Shapiro and Stiglitz [1984], footnote 9):

$$Y = F(L) \equiv \max_{L_i} \sum_{i=1}^{M} f_i(L_i) = \max_{L_i} \{M \cdot f(L_i)\}$$

where by symmetry $L_i = \frac{L}{M}$. Most strikingly, the model is missing an analysis of competition in the sense of how M is determined and who stands behind these firms. This raises the question why the economy does not end in one of the extreme states: either there might be just one producer or each laborer might be self-employed. Recall that all producers are identical and that production function has the standard property of decreasing returns of scale, then the following inequality holds:

$$M \cdot f\left(\frac{L}{M}\right) > f\left(M \cdot \frac{L}{M}\right).$$

This means that the latter case where firms are dissolved completely should be preferable.

[13] To avoid misunderstanding, we do not claim that due to imperfect monitoring the wages offered by firms cannot be higher than what would be the case under perfect monitoring. We only claim that those unemployed are unemployed voluntarily and that the distinction between the NSC curve and the labor supply curve is erroneous.

In order to resolve this issue, we must consider yet another confusing point made by Shapiro and Stiglitz. They dispute the efficiency of market equilibrium by stating, "The natural unemployment rate is too high."[14]

They find that workers are paid their marginal product MPL, and claim that higher employment could be financed from the producer surplus of average product APL over MPL (see the feasibility constraint in equation 12, Shapiro and Stiglitz [1984]). Most importantly Shapiro and Stiglitz do not offer any reason why producer surplus should be redistributed: they simply claim that it belongs to workers.[15]

There is a stark contradiction in their argument: On the one hand, Shapiro and Stiglitz claim that the whole product belongs to workers. On the other hand, they insist on workers being employed not on their own but *in firms*. The solution can be seen easily: If Shapiro and Stiglitz want workers to earn a wage equal to APL, workers should become self-employed and cash in this benefit. Otherwise, if existence of M firms is assumed—as is indeed the case in real economies—there exists a certain value added in firms which naturally belongs to the organizer of the firm.[16]

3 GENERALIZATION: ALTERNATIVE MICRO-ECONOMIC APPROACH

3.1 Deeper Problems of Homogeneous Markets

We have shown that modeling labor as heterogeneous when efficiency is studied results in fundamentally different conclusions

[14] Shapiro and Stiglitz (1984), ibid., p. 440.

[15] This is in effect a direct attack on the marginal revolution in economics, which surprisingly did not receive much attention in subsequent literature. We note that Shapiro's and Stiglitz's proposed solution of redistribution of the surplus APL over MPL is based on nothing else than Karl Marx's surplus value theory.

[16] This value added lies especially in investing the time necessary for the production process, gathering required fixed investments (capital) and managing highly advanced processes of production. These complicated procedures are called *roundabout methods of production* and their distinctive feature lies in investing funds now to consume their products later, instead of enjoying the funds for immediate consumption.

as compared to the single-market approach practised in Solow and Shapiro-Stiglitz models. In this part, we suggest reasons why the heterogeneous labor approach should be strongly preferred. As the reference model, we take the basic Solow (1979) model from section 2.1, but all of our arguments below apply just as well to the Shapiro-Stiglitz model.

The goal of the demand-supply analysis in economics is to study the relationship between prices and quantities. A precondition for this to be possible is to define precisely the good or service whose quantity and price is to be studied. Solow's model, however, violates this requirement. The idea of the model is that higher wages increase workers' effort, for which reason firms may choose to pay above-equilibrium wages and thus create involuntary unemployment. Yet, this very idea implies that by offering higher wages, firms demand *different* type of labor than with lower wages. This means, however, that labor L on the horizontal axis changes its own definition while wage changes. This is a situation that standard microeconomic theory does not allow. The consequences are serious—comparative statics cannot be conducted.

For instance, to compare the amount of employment with the Walrasian equilibrium wage w_{eq} and with Solow model's optimal wage w^*, we would want to know the difference L_{eq}-L^*. Yet, we cannot subtract these two numbers since each of them stands for labor of different quality.[17] This point cannot be overemphasized. In fact, the varying definition of L on the horizontal axis even prevents the use of demand and supply curves. The entire demand-supply

[17] Take a parallel with the firm buying capital equipment—say, trucks. It can buy an expensive, top-quality truck from Mercedes-Benz, or it can buy a cheap, low-quality truck from Daihatsu. However, to subtract the number of Mercedes's from the number of Daihatsus bought does not make any sense. The two trucks have different characteristics despite both of them being trucks. By the same token, it makes little sense to subtract the number of high-quality, expensive bricklayers from the number of low-quality, cheap bricklayers. Certainly, the difference is that one and the same bricklayer can exert different efforts and thus be a high-quality bricklayer at one time and a low-quality one at another. This surely does not hold for trucks. However, this is still not a reason for modeling this situation as one labor market. Markets are abstract concepts defined according to characteristics of goods, not according to the physical fact that two types of goods can be produced by one and the same person. The fact that one single person can do different jobs or exert different efforts is economically irrelevant. What matters is the characteristic of the resulting goods or services.

analysis is built on the assumption of ceteris paribus. If labor is being continually *redefined* when wage is changing, then the ceteris paribus condition is violated: what changes is not only quantity of labor, but also its quality—i.e., the very definition of labor itself. We borrow a statement from Rothbard: "The definition of a *good* is that it consists of an interchangeable supply of one or more units. Therefore, every unit will always be valued equally with every other."[18]

The endogeneity of labor quality on wage causes Solow's model to violate this condition. As a result, it cannot use the standard demand-supply analysis and thus cannot claim to demonstrate involuntary unemployment. In fact, Solow's theory does not model any meaningful market.[19]

In conclusion, if labor is heterogeneous due to model assumptions (effort considerations), then to model it in one homogeneous labor market is not only self-contradictory, but it makes impossible the entire demand-supply analysis that one needs to derive results about market equilibrium. We suggest that the homogeneous labor approach should be replaced by multiple heterogeneous labor markets, differing by the quality of labor offered on them.[20] This is the direction

[18] Rothbard (2004 [1962]), p. 320, emphasis original.

[19] Later on, in a presidential address delivered for the American Economic Association, Solow himself called for a heterogeneous approach to labor markets: "(...) That fact of life merely reminds us that 'labor' is not a well-defined homogenenous factor of production." Solow (1980), p. 4. Yet Solow never changed his 1979 efficiency wage model to incorporate this view. We have shown that this statement of Solow effectively invalidates the conclusions of his own efficiency wage theory.

[20] It is interesting to note that the term efficiency wages was not born in the rich literature of the 1970s and 1980s that claimed the existence of involuntary unemplyoment. Already Alfred Marshall (1964 [1920], p. 456) used the term "efficiency-wages" to explain that wages in the same profession tend to converge to one value (i.e. confirm the law of one price) *after efficiency of the worker is taken into account*. He stated: "Competition tends to make weekly wages in similar employments not equal, but proportionate to the efficiency of the workers." Marshall (1964 [1920]), p. 454. Thus Marshall realized that workers within the same profession differ in efficiency of their work. And he went further: "... it is not an uncommon saying (...) that he is the best business man who contrives to pay the highest wages." Marshall (1964 [1920]), p. 457. This is a strong indication that Marshall, some 60 years before efficiency wage literature appeared, understood that higher wages bring higher effort and thus higher efficiency of labor. But apparently, he saw different efficiencies as different markets. This is why, unlike Solow, Shapiro or Stiglitz, he made no mention of involuntary unemployment.

of current research, among existing papers we refer to research by Bagger, Christensen and Mortensen (2010). In the next sections, we offer a formal setting for the heterogeneous market approach.

3.2 Hedonic Wage Models

One approach that incorporates heterogeneity in the labor market is to consider hedonic wage equations which are commonly integrated in the class of search-matching models. Hedonic wages postulate that workers derive different utility from different jobs, given working conditions and other nonmonetary benefits.

Taking the version presented by Lucas (1977), the overall model starts from a very general description of the market. On the supply side, utility of workers from a job is a function not only of wage, but also of both job characteristics (specific to each job) and personal characteristics (specific to each worker), as in Lucas (1977, eq. 2). In the same way on the demand side, utility of firms from a worker is a function not only of wage, but also of both job characteristics and personal characteristics, as in Lucas (1977, eq. 4). Equilibrium is then defined by equating the supply and demand equations.

This model was novel in that it acknowledged the heterogeneity on both sides of the market. Unfortunately, it appears that its potential was not fully exploited and that it subsequently led to further confusion. For one, its generality does not leave much space for any noticeable conclusions. For two, it overly emphasizes the supply side, namely the utility of workers from various jobs. Lucas himself (1977, fig. 2) investigates how wages depend on *job* characteristics, which implicitly assumes that the jobs with desired characteristics are there for workers to choose from.[21]

Yet, the emphasis on the supply side did little—and could not do much—to clarify the problems inherent in efficiency wage models. In other words, the issue in discussions about the nature of unemployment is not the utility of workers derived from jobs, but the utility of firms derived from workers. That is, how the

[21] To illustrate our point, consider the recent model by Hwang *et al*. (1998) where workers have identical preferences and productivity but where they choose from different jobs offered by firms.

wage depends on *worker* characteristics. Hence we see that by making workers the principal choice agents and by highlighting job heterogeneity at the expense of worker heterogeneity, hedonic wage models led labor market economists astray.

3.3 The Characteristics Model

3.3.1 Lancaster's Model

To appropriately account for the heterogeneity of workers, we propose to employ the model developed by Lancaster (1966, 1971).

In standard consumer theory, preferences are defined on the space of goods, i.e. consumer utility function takes a vector of goods as its argument. In contrast, Lancaster postulates that consumers in fact choose between goods according to their characteristics, i.e. consumer utility function takes a vector of characteristics as its argument. The link between goods and characteristics is established by means of two axiomatic principles stemming from everyday experience:

1. A single good may possess several different valuable characteristics.
2. A number of different goods may have some of their characteristics exactly the same.

Using mathematical transformation from the goods space to characteristics space, Lancaster (1971) shows several interesting results. If there are fewer characteristics than there are goods, some of the goods will not be purchased at all by a given consumer. They will be those goods which do not have an attractive combination of price-discounted characteristics so as to satisfy the given preferences. Lancaster explains that corner solutions to the optimization problem emerge more often than in textbook consumer theory. Another important property of the characteristics framework is that small changes in prices of goods that are not consumed do not affect equilibrium choice.

Below we show that both properties have important implications when we approach the labor market in a similar way.[22]

[22] It is useful to mention that when it comes to recognizing the importance of product heterogeneity, empirical economics seems to be one step ahead of theory.

3.3.2 Market for Labor Characteristics

It is both a straightforward and, as far as we know, hitherto unexplored idea to extend Lancaster's approach to the labor market, which we propose in line with our analysis in the previous sections.

In our opinion, the characteristics model easily accommodates the main features of labor markets which can be observed in practice, above all, the advanced level of workers' specialization and the elaborate screening during the job matching process for the majority of vacancies. In addition, this model is perfectly appropriate because unlike goods, workers can independently alter the characteristics that they offer (e.g. low or high effort).

Instead of goods we consider laborer *characteristic-types* who offer their work on the market, and each of the workers has a number of characteristics. Again it seems natural to assume that the number of characteristics is lower than the number of individual laborers, since firms generally look for a limited bundle of knowledge and skills.

The two important properties of this model described in section 3.3.1 apply. Given workers' characteristics, firms will not consider all candidates for their vacancies but rather search for those with the most favorable combination of desired characteristics discounted by their prices.

The crucial conclusion of the model concerns the importance of relative prices of labor characteristics: If a low-effort or low-skilled worker is unemployed, either he must lower his wage significantly, or he must change the characteristic offered to high effort or high

---

The characteristics model has been applied in frequently cited empirical studies on demand systems for differentiated products. Perhaps the most influential is the paper by Berry, Levinsohn, and Pakes (BLP, 1995) who estimated demand parameters in the U.S. automobile market.

In order to make the concept of characteristics operational and econometrically sound, BLP worked with observed and unobserved characteristics of cars and estimated the respective demand slopes for the observed (or better: measured) ones. BLP enhanced their model by second choice data (BLP, 2004). Other papers include Bresnahan, Stern and Trajtenberg (BST, 1997), who estimate the demand system for personal computers. These studies demonstrate that heterogeneity of products is a relevant and recognized fact in empirical economics, but much less so in theoretical economics.

skill. Another option for the worker is to look for an employer with a different utility function.

The possibility of involuntary unemployment is excluded, since unemployment stems simply from wage-characteristic mismatch. As long as workers are willing to adjust, the unemployment gap can easily be bridged.

4 CONCLUSIONS

The theory of "efficiency wages" is a popular argument used to show that rational market interaction generates downward wage rigidity and involuntary unemployment. We identify two crucial fallacies of this idea. For one, efficiency wage models ignore the most important feature of labor markets: labor heterogeneity. Thus high-effort and low-effort units of labor are viewed as completely interchangeable, which however from the viewpoint of the firm they surely cannot be. For two, efficiency wage papers are plagued with a significant inner inconsistency: The resulting unemployment is classified as involuntary although it results from the effort decision of the worker where there is nothing rigid other than the worker's very own preferences.

To remove these contradictions we reformulate the theory in a more accurate way. Each of the effort or quality levels has to be viewed as a separate market which has its own clearing quantity and price. As such unemployment is a result of workers' reluctance to adjust to the prevailing market conditions on the respective narrow labor sub-market.

Alternatively, the labor market can be modeled by means of characteristics possessed by workers, as was originally proposed for the goods market by Lancaster (1966). This approach is based on a more complex analysis yet leads to the same conclusions. When some worker is unemployed he finds himself in a corner solution of the market optimization. As a result he has to either significantly alter his wage or significantly alter the bundle of provided characteristics. This way he can turn the corner solution to his side and become employed.

Both of these modifications reinforce our central claim that free market interaction cannot lead to unemployment other than voluntary.

AN APPLICATION: DID HENRY FORD PAY EFFICIENCY WAGES?

In a well-known empirical paper, Raff and Summers (1987) ask whether efficiency wage theories can be applied to notable historical episodes of wage increases. As an exemplary case, they take Henry Ford's automobile factory in Detroit in years before World War I. They conclude that Ford's move to increase wages significantly and the subsequent job queues confirm the existence of involuntary umemployment.

Ford Motor Company saw skyrocketing growth of sales and high profitability since its foundation in 1903. However, mass production based on standardized, repetitive work in unpleasant factory conditions gradually caused dissatisfaction on workers' part. As Raff and Summers document,

> ...worker dissatisfaction took visible form. In 1913, annual turnover at the Ford plant reached 370 percent. Ford had to hire 50,448 men during the course of the year in order to maintain the average labor force at 13,623. (...) At the same time that turnover became so alarming, Ford also faced an epidemic of absenteeism. In 1913, the company suffered a 10 percent daily absenteeism rate.[23]

Ford even complained about workers' "soldiering," i.e., a deliberate (but hidden) boycott of production without declaration of an official strike. The problems became so grave that Ford first responded by increasing wages across-the-board by 15 percent in October 1913. This move helped only temporarily and problems of similar nature returned very shortly. Hence, in January 1914, Ford ordered a wage increase of more than 100 percent, raising the daily wage from USD 2.34 to USD 5.00. The measure was accompanied by a reduction of working hours from 9 to 8 hours a day.

Using this example, Raff and Summers aim to confirm the two major assertions of efficiency wage theories: (a) that the productivity improvement at Ford's plant more than offset higher wages, thereby leading to increased profits, and (b) that the higher wages caused disequilibrium on the labor market, creating involuntary unemployment.

---

[23] Raff and Summers (1987), pp. 63–64.

As regards the first claim, Raff and Summers offer plenty of evidence that the wage increase paid off to Ford. Between 1913 and 1914, the turnover rate fell from 370 percent to 54 percent; the company had to lay off 6 times fewer workers in 1914 than in 1913; better work discipline allowed the company to significantly lower material costs; thanks to higher productivity, it could reduce average workforce by 11 percent. Even with fewer workers and shorter working hours, the plant increased the number of produced cars by 15 percent year-on-year. The overall effect was that Ford's net profit grew, in real terms, by 15 percent between 1913 and 1914.

We do not find this result surprising. Ford found that he may well achieve higher profits with more expensive, but higher-quality labor rather than with cheap, but low-quality labor. As we noted in Section 2.1.5, this is a standard, profit-maximizing choice of inputs, be it labor or capital.

However, Raff and Summers then attempt to prove the second claim, i.e. that the new, higher wage was a disequilibrium one. In their view, this is demonstrated by the job queues that appeared in front of Ford's factory. They quote colorful newspaper reports from the days following the pay rise:

> Twelve thousand men, more than congregated around the plant on any other day last week celebrated the [five-dollar day] with a rush on the plant which resulted in a riot and turning of a fire hose on the crowd in weather but little different from zero.[24]

They further document that the wage increase was not meant to attract more highly skilled workers, given the fact that production techniques at Ford's factory did not require more qualified labor than before. Altogether, these considerations lead Raff and Summers to conclude that "Ford's wage surely exceeded his workers' opportunity cost and put him in the position of rationing jobs." (p. 83).

In our opinion, Raff and Summers fall in the trap of defining types of labor too narrowly. In Section 2.1.4 we made the case that labor must be understood fully heterogeneously, taking account

[24] Raff and Summers (1987), p. 73, brackets original.

of all possible differences. In contrast, Raff and Summers only see differences in types of labor when it comes to qualification and skills. They assert that Ford was not looking for different type of labor when he raised wages because he still demanded the same qualifications and skills. Hence, they see an above-equilibrium wage on the job market after the pay rise. On the contrary, we suggest that Ford in fact *was* looking for a different type of labor because he wanted more reliable, more disciplined and more vigorous labor. These characteristics are no less a "type" of labor than formal skill or qualification. Therefore, Ford went to another market to hire labor, and paid the correspondingly higher wage to obtain it.[25]

The fact that Ford decided to raise wages dramatically after serious problems with absenteeism, boycotts, low morale and poor discipline only corroborates our case. Ford could no longer afford to pay a low-quality input, and decided instead to pay a better one. As we stressed in Section 3.1, when one increases his effort, discipline and reliability in response to higher wage, we are moving from one market to another because the definition of labor L is changing. Therefore, involuntary unemployment does not follow.

The large job queues outside Ford's plant do not make the case for involuntary unemployment either. The fact that a number of applicants show up for a job recruitment is not a proof that all of the applicants have the characteristics that the firm is looking for. Our application of Lancaster's model (Section 3.3.2) shows that unemployment stems from wage-characteristic mismatch. It is not surprising that after the announcement of a 100 percent pay rise, thousands of candidates came to apply for jobs. Yet, Ford was looking for a certain combination of characteristics that the recruitment event was to reveal, and surely only some applicants had them. In other words, it is only *during* the recruitment process—not before it—that the applicant finds out what exact labor market the firm is targeting and whether he fits in it. Therefore, the mere fact that there are more applicants than the firm is willing to take

[25] As we pointed out in section 2.1.4, it is crucial to understand that different *types* of labor may well be represented by the very same laborer. Thus Ford did not necessarily have to hire different laborers, but just had to induce them to exert higher effort.

on can never prove the existence of involuntary unemployment.[26]

REFERENCES

Bagger, Jesper, Bent J. Christensen, and Dale T. Mortensen, 2010. "Wage and Productivity Dispersion: Labor Quality or Rent Sharing." Working paper presented at the ISEO Summer School, June 2011, Iseo, Italy.

Bellante, Don. 1994. "Sticky Wages, Efficiency Wages, and Market Processes." *Review of Austrian Economics* 8, no. 1: 21–23.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63, no. 4: 841–890.

——. 2004. "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market." *Journal of Political Economy* 112, no. 1: 68–105.

Bresnahan, Timothy F., Scott Stern, and Manuel Trajtenberg. 1997. "Market Segmentation and the Sources of Rents from Innovation: Personal Computers in the Late 1980s." *RAND Journal of Economics* 28, no. 1: 17–44.

Carmichael, Lorne. 1985. "Can Unemployment Be Involuntary? Comment." *American Economic Review* 75, no. 5: 1213–1214.

Hwang, Hae-shin, Dale T. Mortensen, and Robert W. Reed. 1998. "Hedonic Wages and Labor Market Search." *Journal of Labor Economics* 16, no. 4: 815–847.

Lancaster, Kelvin J. 1966. "A New Approach to Consumer Theory." *Journal of Political Economy* 74, no. 2: 132–157.

——. 1971. *Consumer Demand—A New Approach*. New York: Columbia University Press.

Lucas, Robert. 1977. "Hedonic Wage Equations and Psychic Wages in the

[26] We find it peculiar that Raff and Summers could use the observation of a job queue as a proof of involuntary unemployment. Anyone who has ever applied for a job in a publicly announced recruitment would confirm that there are often more applicants than the number of posts offered. However, few people would call it a proof of the existence of involuntary unemployment, because during the job interview many applicants are revealed to be *outside* the targeted market. Raff and Summers completely disregard the fact that recruitment is a matching process which reveals the heterogeneity of the labor market.

Returns to Schooling." *American Economic Review* 67, no. 4: 549–558.

Marshall, Alfred. 1920. *Principles of Economics*. Macmillan, London. 1964.

Raff, Daniel M. G., and Lawrence H. Summers. 1987. "Did Henry Ford Pay Efficiency Wages?" *Journal of Labor Economics* 5, no. 4: 57–86. Part 2: The New Economics of Personnel.

Romer, David. 2006. *Advanced Macroeconomics. 3rd ed*. McGraw-Hill. New York.

Rothbard, Murray N. 1962. *Man, Economy, and State*. Auburn, Ala.: Ludwig von Mises Institute. 2004.

Shapiro, Carl, and Joseph E. Stiglitz. 1984. "Equilibrium Unemployment As a Worker Discipline Device." *American Economic Review* 74, no. 3: 433–444.

——. 1985. "Can Unemployment Be Involuntary? Reply." *American Economic Review* 75, no. 5: 1215–1217.

Solow, Robert M. 1979. "Another Possible Source of Wage Stickiness." *Journal of Macroeconomics* 1, no. 1: 79–82.

——. 1980. "On Theories of Unemployment." *American Economic Review* 70, no. 1: 1–11.

Summers, Lawrence H. 1988. "Relative Wages, Efficiency Wages, and Keynesian Unemployment." *American Economic Review* 78, no. 2: 383–388.

Yellen, Janet L. 1984. "Efficiency Wage Models of Unemployment." *American Economic Review* 74, no. 2: 200–205.