Economic Quarterly—Volume 94, Number 3—Summer 2008—Pages 197–218

New Keynesian Economics: A Monetary Perspective

Stephen D. Williamson

Since John Maynard Keynes wrote the *General Theory of Employment, Interest, and Money* in 1936, Keynesian economics has been highly influential among academics and policymakers. Keynes has certainly had his detractors, though, with the most influential being Milton Friedman, Robert Lucas, and Edward C. Prescott. Monetarist thought, the desire for stronger theoretical foundations in macroeconomics, and real business cycle theory have at times been at odds with Keynesian economics. However, Keynesianism has remained a strong force, in part because its practitioners periodically adapt by absorbing the views of its detractors into the latest "synthesis."

John Hicks's IS-LM interpretation of Keynes (Hicks 1937) and the popularization of this approach, particularly in Samuelson's textbook (Samuelson 1997), gave birth to the "neoclassical synthesis." Later, the menu cost models developed in the 1980s were a response to a drive for a more serious theory of sticky prices (Mankiw 1985, Caplin and Spulber 1987). More recently, New Keynesian economists have attempted to absorb real business cycle analysis and other ideas from post-1972 macroeconomics into a "new neoclassical synthesis" (Goodfriend and King 1997).

The important New Keynesian ideas, as summarized, for example in Clarida, Galí, and Gertler (1999) and Woodford (2003), are the following:

1. The key friction that gives rise to short-run nonneutralities of money and the primary concern of monetary policy is sticky prices. Because some prices are not fully flexible, inflation or deflation induces relative price distortions and welfare losses.

Williamson is affiliated with Washington University in St. Louis, the Federal Reserve Bank of Richmond, and the Federal Reserve Bank of St. Louis. He thanks Huberto Ennis, Robert Hetzel, John Weinberg, and Alex Wolman for helpful comments and suggestions. E-mail: swilliam@artsci.wustl.edu.

2. Modern monetary economics is not part of the New Keynesian synthesis. New Keynesians typically regard the frictions that we encounter in deep (e.g., Lagos and Wright 2005) and not-so-deep (e.g., Lucas and Stokey 1987) monetary economics as being second-order importance. These frictions are absence-of-double-coincidence problems and information frictions that give rise to a fundamental role for monetary exchange, and typically lead to intertemporal distortions that can be corrected by monetary policy (for example, a ubiquitous result in monetary economics is Friedman's zero-nominal-interest-rate rule for correcting intertemporal monetary distortions). The Friedman rule is certainly not ubiquitous in New Keynesian economics.

3. The central bank is viewed as being able to set a short-term nominal interest rate, and the monetary policy problem is presented as the choice over alternative rules for how this nominal interest rate should be set in response to endogenous and exogenous variables.

4. There is a short-run Phillips curve tradeoff. A monetary policy that produces an increase in unanticipated inflation will tend to increase real aggregate output.

The goal of this paper is to construct a simple sticky-price New Keynesian model and then use it to understand and evaluate the ideas above. In this model there are some important departures from the typical New Keynesian models studied by Clarida, Galí, and Gertler; Woodford; and others. However, these departures will highlight where the central ideas and results in New Keynesian analysis are coming from.

For monetary economists, key aspects of New Keynesian economics can be puzzling. For example in Woodford (2003), the apparently preferred framework for analysis is a "cashless model" in which no outside money is held in equilibrium. Prices are denominated in terms of some object called money, and these prices are assumed to be sticky. The interest rate on a nominal bond can be determined in the cashless model, and the central bank is assumed capable of setting this nominal interest rate. Then, the monetary policy problem is formulated as the choice over rules for setting this nominal interest rate. This approach can be contrasted with the common practice in monetary economics, where we start with a framework in which money overcomes some friction, serves as a medium of exchange, and is held in equilibrium in spite of being dominated in rate of return by other assets. Then, studying the effects of monetary policy amounts to examining the consequences of changing the stock of outside money through various means: open market operations, central bank lending, or outright "helicopter drops." It is usually possible to consider monetary policy rules that dictate the contingent behavior of a nominal interest rate, but in most monetary models we can see what underlying actions the central bank must take concerning monetary quantities to support such a

policy. What is going on here? Is the New Keynesian approach inconsistent with the principles of monetary economics? Is it misleading?

The first task in this article is to construct a cashless model with sticky prices. This model departs from the usual New Keynesian construct in that there are competitive markets rather than Dixit-Stiglitz (1977) monopolistic competition. This departure helps to make the model simple, and yields information on the importance of the noncompetitive behavior of firms for New Keynesian economics. In general, given the emphasis on the sticky-price friction in New Keynesian economics, we would hope that it is something inherent in the functioning of a sticky-price economy, rather than simply strategic behavior, that is at the heart of the New Keynesian mechanism.

Our cashless model is consistent with most of the predictions of standard Keynesian models, new and old. The sticky-price friction leads to a relative price distortion in that with inflation (deflation), too large (too small) a quantity of sticky-price goods is produced and consumed relative to flexible-price goods. Optimally, the inflation rate is zero, which eliminates the relative price distortion. One aspect in which this model differs from standard New Keynesian models is that it does not exhibit Phillips curve correlations. If the substitution effect dominates in the labor supply response to a wage increase (which we consider the standard case), then output is decreasing (increasing) in the inflation rate when the inflation rate is positive (negative). This is because the distortion caused by sticky prices rises with the deviation from a constant price level, and the representative consumer supplies less labor in response to a larger sticky-price distortion. Thus, under these circumstances output is maximized when the inflation rate is zero and there is no output/inflation tradeoff. In the case where the income effect dominates in the labor supply response to a change in wages, output increases (decreases) for positive (negative) inflation rates. Here there is a Phillips curve tradeoff if the inflation rate is positive, but a zero inflation rate is optimal.

In most New Keynesian models, Phillips curve correlations are generated because of the strategic forward-looking behavior of price-setting firms. A firm, given the opportunity to set the price in units of money for its product, knows it will not have this opportunity again until some time in the future. Roughly, what matters to the firm is its expectation of the path for the price level during the period of time until its next price-setting opportunity. If the inflation rate is unusually high in the future, then the firm's relative price will be unexpectedly low, and then, by assumption, it will be satisfying higher-than-expected demand for its product. Given that all firms behave in the same way, unanticipated inflation will tend to be associated with high real output. One message from our model is that Phillips curve behavior can disappear in the absence of strategic behavior by firms, even with sticky prices.

We live in a world where outside money is held by consumers, firms, and financial institutions in the form of currency and reserve balances, and this

outside money is supplied by the central bank and used in various transactions at the retail level and among financial institutions. In using a cashless model to analyze monetary policy, we should feel confident that we are not being led astray by a quest for simplicity. To evaluate what we might lose in focusing on a cashless model, we develop a monetary model that is a straightforward cash-in-advance extension of the cashless model. Then, in the spirit of Woodford (2003), we explore how the behavior of this model compares to that of the cashless model and study the "cashless limit" of the more elaborate model. As it turns out, it requires very special assumptions for the limiting economy to behave in the same way as the monetary economy, and, in any case, quantity theory principles hold in equilibrium in the monetary economy. It is useful to know how monetary quantities should be manipulated to produce particular time paths of nominal interest rates, prices, and quantities in the economy, as the key instruments that a central bank has available to it are the quantities on its balance sheet. Thus, it would seem preferable to analyze monetary economies rather than cashless economies.

This article is organized as follows. In Section 1 we construct the basic cashless model, work through examples, and uncover the general properties of this model. Section 2 contains a monetary model, extending the cashless model as a cash-in-advance construct with money and credit. Section 3 is a detailed discussion of the importance of the results, and Section 4 is a conclusion.

1. CASHLESS MODEL

The goal of this section of the paper is to construct a simple sticky-price model that will capture the key aspects of New Keynesian economics, while also taking a somewhat different approach to price determination, in order to simplify and illuminate the important principles at work. The model we construct shares features with typical New Keynesian "cashless" models (see Woodford 2003), which are the following:

1. Money is not useful for overcoming frictions, it does not enter a cash-in-advance constraint or a utility function, nor does it economize on transactions costs.

2. Money is a numeraire in which all prices are denominated.

3. The money prices of goods are sticky in that, during any period, some goods prices are predetermined and do not respond to current aggregate shocks.

This model captures the essential friction that New Keynesians argue should be the focus of monetary policy—a sticky-price friction. New Keynesians argue that the other frictions that we typically encounter in monetary models—absence-of-double-coincidence problems and intertemporal

price distortions, for example—are of second order for the problem at hand. Further, New Keynesians feel that it is important to model the monetary policy problem in terms of the choice of nominal interest rate rules and that this framework is a very convenient vehicle in that respect.

The model we will work with here has an infinite-lived representative consumer who maximizes

$$E_0 \sum_{t=0}^{\infty} \beta^t \left[u(c_t^1) + u(c_t^2) - v(n_t) \right], \tag{1}$$

where c_t^i denotes consumption of the i^{th} good, $i = 1, 2$, and n_t is labor supply. Assume that $u(\cdot)$ is strictly increasing, strictly concave, twice continuously differentiable, and has the property $u'(0) = \infty$. As well, $v(\cdot)$ is strictly increasing, strictly convex, and twice differentiable with $v'(0) = 0$ and $v'(h) = \infty$ for some $h > 0$. We have assumed a separable period utility function for convenience, and a two-good model is sufficient to exposit the ideas of interest. Goods are perishable. There are linear technologies for producing goods from labor input, i.e.,

$$y_t^i = \gamma_t n_t^i, \tag{2}$$

where y_t^i is output of good i, γ_t is aggregate productivity, and n_t^i is the quantity of labor input applied to production of good i. Assume that γ_t follows an exogenous stochastic process.

This model is very simple. There is no investment or capital, and an optimal allocation is a sequence $\{\tilde{n}_t^1, \tilde{n}_t^2, \tilde{c}_t^1, \tilde{c}_t^2\}_{t=0}^{\infty}$ satisfying $\tilde{n}_t^1 = \tilde{n}_t^2 = \tilde{n}_t$ and $\tilde{c}_t^1 = \tilde{c}_t^2 = \tilde{c}_t$, with $\tilde{c}_t = \gamma_t \tilde{n}_t$, where $\tilde{n}_t$ solves

$$\frac{v'(2\tilde{n}_t)}{u'(\gamma_t \tilde{n}_t)} = \gamma_t. \tag{3}$$

Therefore, at the optimum, consumption of the two goods should be equal in each period with the same quantity of labor allocated to production of each good, given symmetry. As well, from (3) the ratio of the marginal disutility of labor to the marginal utility of consumption should be equal to aggregate productivity for each good in each period.

There is another object, which we will call money, that plays only the role of numeraire. We will assume a form of price stickiness reminiscent of what obtains in the staggered wage-setting models of Fischer (1977) and Taylor (1979). That is, assume that prices are sticky, in the sense that, if the price of a good is flexible in period t, then it remains fixed at its period t value through period $t + 1$, and is subsequently flexible in period $t + 2$, etc. In any given period, one good is flexible and the other is sticky. Now, since in (1) and (2) the two goods are treated symmetrically in preferences and technology, we can let good 1 be the flexible-price good in each period. Then, let P_t denote the price in units of money of the flexible-price good in period t, and then P_{t-1}

is the price of the sticky-price good. As well, let W_t denote the nominal wage rate.

The Keynesian modeler must always deal with the problem of how firms and consumers behave in the face of price and/or wage stickiness, as well as how quantities are determined. In typical textbook sticky-wage approaches, the nominal wage is exogenous, and the quantity of labor traded is determined by what is optimal for the representative firm. Standard textbook sticky-price models have a single good sold at an exogenous nominal price, and the quantity of output is demand-determined.[1]

In New Keynesian economics, the emphasis is on price stickiness (as opposed to wage stickiness), the distribution of prices across goods, and relative price distortions. Clearly, if there is a homogeneous good, constant returns to scale, competitive equilibrium, and perfect information, we cannot have anything other than a degenerate distribution of prices in equilibrium, where all firms producing positive output charge the same price. The New Keynesian approach at least requires heterogeneous goods, and the standard model in the literature is currently one with monopolistically competitive firms. For example, a typical approach is to assume monopolistic competition with Calvo (1983) price setting (or "time-dependent pricing"). In such a model, each firm randomly obtains the opportunity to change its nominal price each period, so that with a continuum of firms, some constant fraction of firms changes their prices optimally, while the remaining fraction is constrained to setting prices at the previous period's values. Alternatively, it could be assumed that each of the monopolistically competitive firms must set their prices one period in advance, before observing aggregate (or possibly idiosyncratic) shocks. Woodford (2003), for example, takes both approaches.

Neither Calvo pricing nor price setting one period in advance in monopolistic competition models is without problems. In a Calvo pricing model, each monopolistically competitive firm is constrained to producing one product. There may be states of the world where some firms will earn negative profits, but they are somehow required to produce strictly positive output anyway, and firms cannot reallocate productive factors to the production of flexible-price goods that will yield higher profits. With prices set one period in advance, firms are constrained to produce, even in the face of negative ex post profits.

Here, we have assumed differentiated products, but we will maintain competitive pricing. This model is certainly not typical, so it requires some explanation. First, the consumer's budget constraint will imply that all wage income will be spent on the two consumption goods, so

$$P_t\gamma_t n_t^1 + P_{t-1}\gamma_t n_t^2 - W_t(n_t^1 + n_t^2) = 0. \tag{4}$$

[1] See Williamson (2008) for examples of standard Keynesian sticky-wage and sticky-price models.

However, this would then seem to imply that, unless $P_t = P_{t-1}$, the production and sale of one good will earn strictly positive profits and production and sale of the other good will yield strictly negative profits. Thus, it seems that it cannot be profit-maximizing for both goods to be produced in equilibrium. However, suppose that we take for granted, as is typical in much of the New Keynesian literature, that there will be some firms that are constrained to producing the sticky-price good, and that the firms who produce this good will satisfy whatever demand arises at the price P_{t-1}. How then should we determine which firms produce which good? For this purpose, assume that there is a lottery, which works as follows. Before a firm produces, it enters a lottery where the outcome of the lottery determines whether the firm produces the flexible-price good or the sticky-price good. If it is determined through the lottery that a particular firm produces the flexible-price good, then that firm receives a subsidy of s_t^1 in nominal terms, per unit of output. If it is determined that a particular firm produces the fixed-price good, that firm receives s_t^2 per unit of output. The agent that offers the lottery will set s_t^1 and s_t^2 so that any firm is indifferent between producing the fixed-price and flexible-price goods, i.e.,

$$P_t - s_t^1 = P_{t-1} - s_t^2. \tag{5}$$

Further, the agent offering the lottery breaks even, so that

$$s_t^1 n_t^1 + s_t^2 n_t^2 = 0, \tag{6}$$

so solving for the subsidy rates, we obtain

$$s_t^1 = \frac{n_t^2 (P_t - P_{t-1})}{(n_t^1 + n_t^2)} \text{ and}$$

$$s_t^2 = \frac{n_t^1 (P_{t-1} - P_t)}{(n_t^1 + n_t^2)}.$$

Given these subsidy rates, the agent offering the lottery breaks even and each firm is willing to enter the lottery as profits per unit produced are zero in equilibrium, whether the firm ultimately produces the flexible-price or sticky-price good. Though this cross-subsidization setup may seem unrealistic, it does not seem less palatable than what occurs in typical Keynesian sticky-price models. In fact, the randomness in determining which firms produce which good is reminiscent of the randomness in Calvo (1983) pricing, but our approach is much more tractable.

Now, let π_t denote the relative price of flexible-price and sticky-price goods, which is also the gross rate of increase in the price of the flexible-price good. Under some special circumstances, π_t will also be the measured inflation rate, but in general that is not the case. However, π_t is the relative price that captures the extent of the effects of the sticky-price friction. Letting

w_t denote the relative price of labor and flexible-price goods, we can rewrite equation (4) as

$$\gamma_t n_t^1 + \frac{\gamma_t n_t^2}{\pi_t} - w_t(n_t^1 + n_t^2) = 0. \tag{7}$$

Optimization by the consumer is summarized by the following two marginal conditions:

$$u'(c_t^1) - \pi_t u'(c_t^2) = 0 \text{ and} \tag{8}$$

$$w_t u'(c_t^1) - v'(n_t^1 + n_t^2) = 0. \tag{9}$$

In equilibrium all output must be consumed, so

$$c_t^i = \gamma_t n_t^i \tag{10}$$

for $i = 1, 2$. Further, letting q_t denote the price at time t of a claim to one unit of money in period $t + 1$, we can determine q_t in the usual fashion by

$$q_t = \beta E_t \left[\frac{u'(c_{t+1}^1)}{\pi_{t+1} u'(c_t^1)} \right]. \tag{11}$$

An equilibrium is defined to be a stochastic process $\{\pi_t, w_t, q_t, n_t^1, n_t^2, c_t^1, c_t^2\}_{t=0}^{\infty}$ given an exogenous stochastic process for γ_t, that satisfies (7)–(11) and $q_t \leq 1$, so that the nominal interest rate is nonnegative in each state of the world. There is clearly indeterminacy here, as there appears to be nothing that will pin down prices. In equilibrium there is an object called money that is in zero supply, and which, for some unspecified reason, serves as a unit of account in which prices are denominated. The path that π_t follows in equilibrium clearly matters, because prices are sticky, but the possibilities for equilibrium paths for π_t are limitless.

Examples

One equilibrium is $\pi_t = 1$ for all t, which from (7)–(11) gives the optimal allocation with $\tilde{n}_t^1 = \tilde{n}_t^2 = \tilde{n}_t$ and $\tilde{c}_t^1 = \tilde{c}_t^2 = \tilde{c}_t$, where $\tilde{c}_t = \gamma_t \tilde{n}_t$, and where $\tilde{n}_t$ solves (3). Solving for q_t from equation (11), we get

$$q_t = \beta E_t \left[\frac{u'(\gamma_{t+1} \tilde{n}_{t+1})}{u'(\gamma_t \tilde{n}_t)} \right], \tag{12}$$

and so long as the variability in productivity is not too large, we will have $q_t \leq 1$ for all t, so that the nominal interest rate is always nonnegative, and this is indeed an equilibrium.

Alternatively (for example, see Woodford 2003), we could argue that the central bank can set the nominal interest rate $i_t = \frac{1}{q_t} - 1$. In this instance, if the central bank sets the nominal interest rate from equation (12) according to

$$i_t = \left\{ \beta E_t \left[\frac{u'(\gamma_{t+1} \tilde{n}_{t+1})}{u'(\gamma_t \tilde{n}_t)} \right] \right\}^{-1} - 1, \tag{13}$$

an equilibrium with $\pi_t = 1$ for all t can be achieved, which is optimal.

There is nothing in the model that tells us why the central bank can control i_t, and why it cannot control π_t, for example. In New Keynesian models, the justification for treating the market nominal interest rate as a direct instrument of the central bank comes from outside the model, along the lines of "this is what most central banks do." In any case, an optimal policy implies that, since $\pi_t = 1$, the price level is constant and the inflation rate is zero. This optimal policy could then be characterized as an inflation rate peg, or as a policy that requires, from (13), that the nominal interest rate target for the central bank fluctuate with the aggregate technology shock. More simply, from (13), the nominal interest rate at the optimum should equal the "Wicksellian natural rate of interest" (see Woodford 2003).

There are, of course, many suboptimal equilibria in this model. For example, consider the special case where $u(c) = \ln c$ and $v(n) = \delta n$, for $\delta > 0$. Though $v(\cdot)$ does not satisfy some of our initial restrictions, this example proves particularly convenient. We will first construct an equilibrium with a constant inflation rate, which has the property that $\pi_t = \alpha$, where α is a positive constant (recall that π_t is not the gross inflation rate, but if π_t is constant then the inflation rate is constant). From (7)–(11), the equilibrium solution we obtain is

$$w_t = \frac{2\gamma_t}{1+\alpha}, \tag{14}$$

$$q_t = \frac{\beta\gamma_t}{\alpha} E_t\left(\frac{1}{\gamma_{t+1}}\right), \tag{15}$$

$$n_t^1 = \frac{1}{\delta(1+\alpha)}, \tag{16}$$

$$n_t^2 = \frac{\alpha}{\delta(1+\alpha)}, \tag{17}$$

$$c_t^1 = \frac{\gamma_t}{\delta(1+\alpha)}, \text{ and} \tag{18}$$

$$c_t^1 = \frac{\alpha\gamma_t}{\delta(1+\alpha)}. \tag{19}$$

In this equilibrium, the rate of inflation is $\alpha - 1$. Note, from (14)–(19), that higher inflation causes a reallocation of consumption and labor supply from flexible-price to sticky-price goods. From equation (15) for $q_t \leq 1$, it is sufficient that α not be too small and that γ_t not be too variable. This equilibrium is of particular interest because it involves an inflation rate peg. Of course, as we showed previously, $\alpha = 1$ is optimal.

Alternatively, consider the same example as above, with log utility from consumption and linear disutility to supplying labor, but now suppose that π_t is governed by a rule that responds to productivity shocks, for example,

$$\pi_t = \frac{\gamma_{t-1}}{\gamma_t}.$$

Then, from (7)–(11), we obtain the equilibrium solution

$$w_t = \frac{2\gamma_t^2}{\gamma_{t-1} + \gamma_t}, \tag{20}$$

$$q_t = \beta E_t \left[\frac{\gamma_t(\gamma_t + \gamma_{t+1})}{\gamma_{t+1}(\gamma_{t-1} + \gamma_t)} \right], \tag{21}$$

$$n_t^1 = \frac{2\gamma_t}{\delta(\gamma_{t-1} + \gamma_t)}, \tag{22}$$

$$n_t^2 = \frac{2\gamma_{t-1}}{\delta(\gamma_{t-1} + \gamma_t)}, \tag{23}$$

$$c_t^1 = \frac{2\gamma_t^2}{\delta(\gamma_{t-1} + \gamma_t)}, \text{ and} \tag{24}$$

$$c_t^1 = \frac{2\gamma_t\gamma_{t-1}}{\delta(\gamma_{t-1} + \gamma_t)}. \tag{25}$$

From (21) it is sufficient for the existence of equilibrium that γ_t not be too variable. Note in the solution, (20)–(25), that equilibrium quantities and prices all exhibit persistence because of the contingent path that prices follow. Complicated dynamics can be induced through the nominal interest rate rule, of which (25) is an example in this case. Indeed, it is possible (see Woodford 2003), given some nominal interest rate rules, to obtain equilibrium solutions where current endogenous variables depend on anticipated future aggregate shocks. Typical New Keynesian models also obtain equilibrium solutions with such properties through the forward-looking price-setting behavior of monopolistically competitive producers. This latter mechanism is not present in our model.

General Properties of the Model

To further analyze our model, we find it useful to consider how we would solve for an equilibrium in this model in the absence of sticky prices. The model is purely static, so we can solve period-by-period. An equilibrium for period t consists of relative prices π_t and w_t, and quantities n_t^1, n_t^2, c_t^1, and

c_t^2 that solve the marginal conditions (8) and (9), the equilibrium conditions (10), and two zero-profit conditions:

$$\gamma_t n_t^1 - w_t n_t^1 = 0 \text{ and} \tag{26}$$

$$\frac{\gamma_t n_t^2}{\pi_t} - w_t n_t^2 = 0. \tag{27}$$

Of course, the solution we get is the optimum, with $w_t = \gamma_t$, $\pi_t = 1$, $n_t^1 = n_t^2 = \tilde{n}_t$, and $c_t^1 = c_t^2 = \tilde{c}_t = \gamma_t \tilde{n}_t$, with $\tilde{n}_t$ determined as the solution to (3).

Now, how should we think about solving the system (7)–(11) under sticky prices? It seems most useful to think of this system in the traditional Keynesian sense, as a model where one price, π_t, is fixed exogenously. Given any exogenous $\pi_t \neq 1$, it cannot be the case that all agents optimize and all markets clear in equilibrium. The solution we have chosen here, which is in line with standard New Keynesian economics, is to allow for the fact that (26) and (27) do not both hold. Instead, we allow for cross-subsidization with zero net subsidies across production units and zero profits in equilibrium net of subsidies for production of each good, which gives us equation (7).

Given this interpretation of the model, how should we interpret q_t, as determined by equation (11)? Since π_t is simply the relative price of good 2 in terms of good 1 in period t, q_t is the price, in units of good 1 in period t, that a consumer would pay for delivery of one unit of good 2 in period $t+1$. Why, then, should we require an arbitrage condition that $q_t \leq 1$, or why should

$$\beta E_t \left[\frac{u'(c_{t+1}^1)}{\pi_{t+1} u'(c_t^1)} \right] \leq 1 \tag{28}$$

hold? Such a condition requires the existence of a monetary object. Then we can interpret (11) as determining the price in units of money in period t of a claim to one unit of money delivered in period $t+1$, and inequality (28) is required so that zero money balances are held in equilibrium. Thus, in equilibrium the model is purely atemporal. There is no intertemporal trade, nevertheless there exist equilibrium prices for money in terms of goods and for the nominal bond in each period.

Thus far, this may be somewhat puzzling for most monetary economists, who are accustomed to thinking about situations in which money is not held in equilibrium as ones in which the value of money in units of goods is zero in each period. This does not happen here, but no fundamental principles of economic analysis appear to have been violated.

The key question, then, is what we can learn from this model. First, we will get some idea of the operating characteristics of the model through linear approximation. If we treat π_t as exogenous, following our interpretation above, then the exogenous variables are γ_t and π_t. Substitute using equation (10) in (7)–(9), and then linearize around the solution we get with $\gamma_t = \bar{\gamma}$ and

$\pi_t = 1$, where $\bar{\gamma}$ is a positive constant. The equilibrium solution we get in this benchmark case is $w_t = \bar{\gamma}$, $n_t^1 = n_t^2 = \bar{n}$, and $c_t^1 = c_t^2 = \bar{\gamma}\bar{n}$, where $\bar{n}$ solves

$$\bar{\gamma}u'(\bar{\gamma}\bar{n}) - v'(2\bar{n}) = 0.$$

The solution to the linearized model is (leaving out the solution for the wage, w_t) :

$$n_t^1 = \bar{n} + \frac{u'(-2v'' + \bar{\gamma}^2 u'')}{2\bar{\gamma}u''(\bar{\gamma}^2 u'' - 2v'')}(\pi_t - 1) - \frac{u' + \bar{\gamma}\bar{n}u''}{\bar{\gamma}^2 u'' - 2v''}(\gamma_t - \bar{\gamma}), \tag{29}$$

$$n_t^2 = \bar{n} + \frac{u'(2v'' - \bar{\gamma}^2 u'')}{2\bar{\gamma}u''(\bar{\gamma}^2 u'' - 2v'')}(\pi_t - 1) - \frac{u' + \bar{\gamma}\bar{n}u''}{\bar{\gamma}^2 u'' - 2v''}(\gamma_t - \bar{\gamma}), \tag{30}$$

$$c_t^1 = \bar{\gamma}\bar{n} + \frac{u'(-2v'' + \bar{\gamma}^2 u'')}{2u''(\bar{\gamma}^2 u'' - 2v'')}(\pi_t - 1) - \frac{2\bar{n}v'' + \bar{\gamma}u'}{\bar{\gamma}^2 u'' - 2v''}(\gamma_t - \bar{\gamma}), \text{ and} \tag{31}$$

$$c_t^2 = \bar{\gamma}\bar{n} + \frac{u'(2v'' - \bar{\gamma}^2 u'')}{2u''(\bar{\gamma}^2 u'' - 2v'')}(\pi_t - 1) - \frac{2\bar{n}v'' + \bar{\gamma}u'}{\bar{\gamma}^2 u'' - 2v''}(\gamma_t - \bar{\gamma}), \tag{32}$$

and aggregate labor supply and output are given, respectively, by

$$n_t^1 + n_t^2 = 2\bar{n} - \frac{2(u' + \bar{\gamma}\bar{n}u'')}{\bar{\gamma}^2 u'' - 2v''}(\gamma_t - \bar{\gamma}) \text{ and} \tag{33}$$

$$c_t^1 + c_t^2 = 2\bar{\gamma}\bar{n} - \frac{2(2\bar{n}v'' + \bar{\gamma}u')}{\bar{\gamma}^2 u'' - 2v''}(\gamma_t - \bar{\gamma}). \tag{34}$$

Therefore, from (29)–(34), in the neighborhood of an equilibrium with constant prices, an increase in π_t, which corresponds to an increase in the inflation rate, results in a decrease in the production and consumption of the flexible-price good, an increase in production and consumption of the fixed-price good, and no effect on aggregate labor supply and output. Thus, this model does not produce a Phillips curve correlation, at least locally, and increases in the inflation rate serve only to misallocate production and consumption across goods.

As well, from (29)–(34), a positive shock to aggregate productivity has the same effect on production and consumption of both goods. If

$$u' + \bar{\gamma}\bar{n}u'' > 0, \tag{35}$$

then the substitution effect of the productivity increase offsets the income effect on labor supply so that aggregate labor supply increases. In what follows, we assume that the substitution effect dominates, i.e., (35) holds. From (34), aggregate output increases with an increase in productivity, regardless of whether the substitution effect dominates.

The absence of a Phillips curve effect here might seem puzzling, as a positive relationship between inflation and output often appears as a cornerstone of new and old Keynesian economics. Thus, we should explore this further to see how π_t affects output outside of the neighborhood of our baseline equilibrium. Consider an example that will yield closed-form solutions, in particular $u(c) = \frac{c^{1-\alpha}-1}{1-\alpha}$ and $v(n) = \delta n$, with $\alpha > 0$ and $\delta > 0$. Solving (7)–(10), we obtain

$$n_t^1 = \gamma_t^{\frac{1}{\alpha}-1}\delta^{-\frac{1}{\alpha}}\left(\frac{1+\pi_t^{\frac{1}{\alpha}-1}}{1+\pi_t^{\frac{1}{\alpha}}}\right)^{\frac{1}{\alpha}}, \tag{36}$$

$$n_t^2 = \gamma_t^{\frac{1}{\alpha}-1}\delta^{-\frac{1}{\alpha}}\left(\frac{\pi_t^{\frac{1}{\alpha}}+\pi_t^{\frac{2}{\alpha}-1}}{1+\pi_t^{\frac{1}{\alpha}}}\right)^{\frac{1}{\alpha}}, \tag{37}$$

$$c_t^1 = \gamma_t^{\frac{1}{\alpha}}\delta^{-\frac{1}{\alpha}}\left(\frac{1+\pi_t^{\frac{1}{\alpha}-1}}{1+\pi_t^{\frac{1}{\alpha}}}\right)^{\frac{1}{\alpha}}, \tag{38}$$

$$c_t^2 = \gamma_t^{\frac{1}{\alpha}}\delta^{-\frac{1}{\alpha}}\left(\frac{\pi_t^{\frac{1}{\alpha}}+\pi_t^{\frac{2}{\alpha}-1}}{1+\pi_t^{\frac{1}{\alpha}}}\right)^{\frac{1}{\alpha}}, \text{ and} \tag{39}$$

$$w_t = \gamma_t\left(\frac{1+\pi_t^{\frac{1}{\alpha}-1}}{1+\pi_t^{\frac{1}{\alpha}}}\right). \tag{40}$$

Then, aggregate labor supply and aggregate output are given by

$$n_t^1 + n_t^2 = \gamma_t^{\frac{1}{\alpha}-1}\delta^{-\frac{1}{\alpha}}\left(1+\pi_t^{\frac{1}{\alpha}-1}\right)^{\frac{1}{\alpha}}\left(1+\pi_t^{\frac{1}{\alpha}}\right)^{1-\frac{1}{\alpha}} \text{ and} \tag{41}$$

$$c_t^1 + c_t^2 = \gamma_t^{\frac{1}{\alpha}}\delta^{-\frac{1}{\alpha}}\left(1+\pi_t^{\frac{1}{\alpha}-1}\right)^{\frac{1}{\alpha}}\left(1+\pi_t^{\frac{1}{\alpha}}\right)^{1-\frac{1}{\alpha}}. \tag{42}$$

Now, in the solution (36)–(42), the condition (35) is equivalent to $\alpha < 1$. Given this, labor supply in both sectors is increasing in productivity, as, of course, is consumption of each good and total output.

Our primary interest in this example is what it tells us about the relationship between π_t and aggregate output. Note from equation (42) that this relationship is determined by the properties of the function

$$G(\pi) = \left(1+\pi^{\frac{1}{\alpha}-1}\right)^{\frac{1}{\alpha}}\left(1+\pi^{\frac{1}{\alpha}}\right)^{1-\frac{1}{\alpha}}.$$

Differentiating, we obtain

$$G'(\pi) = \pi^{\frac{1}{\alpha}-2}\left(1+\pi^{\frac{1}{\alpha}-1}\right)^{\frac{1}{\alpha}-1}\left(1+\pi^{\frac{1}{\alpha}}\right)^{-\frac{1}{\alpha}}\frac{1}{\alpha}\left(\frac{1}{\alpha}-1\right)(1-\pi).$$

Therefore, for the case $\alpha < 1$, we have $G'(1) = 0$, $G'(\pi) > 0$ for $\pi < 1$, and $G'(\pi) < 0$ for $\pi > 1$. Thus, output is maximized for $\pi = 1$ and the Phillips curve has a negative slope when inflation is positive and a positive slope when inflation is negative. The key to the Phillips curve relationship is how labor supply responds to the distortion created by inflation or deflation due to the the sticky-price friction. When the substitution effect on labor supply of an increase in productivity dominates the income effect, an increase or decrease in the inflation rate from zero implies that the marginal payoff from supplying labor falls, and the consumer therefore reduces labor supply.

The interesting aspect of these results is that they point to a nonrobust link between price stickiness and Phillips curve correlations. In spite of the fact that firms do not set prices strategically in a forward-looking manner, intuition might tell us that there should still be a Phillips curve correlation. That is, with higher inflation, it might seem that the additional quantity of output produced by sticky-price firms should be greater than the reduction in output by flexible-price firms, and aggregate output should increase. However, our analysis shows that this need not be the case and that the key to understanding the mechanism at work is labor supply behavior.

In our model, since not all prices are sticky, the key effect of inflation on output comes from the relative price distortion, and labor supply may increase or decrease in response to higher inflation, with a decrease occurring when the elasticity of substitution of labor supply is sufficiently high. As we commented earlier, the assumptions on price stickiness and firm behavior in our model seem no less palatable than what is typically assumed. Thus, the nonrobustness of the Phillips curve we find here deserves attention.

2. A MONETARY MODEL AND THE "CASHLESS LIMIT"

In New Keynesian economics (e.g., Woodford 2003), baseline "cashless" models are taken seriously as frameworks for monetary policy analysis. As we have seen, the cashless model focuses attention on the sticky-price friction as the key source of short-run nonneutralities of money. New Keynesian arguments for using a cashless model appear to be as follows: (i) the standard intertemporal monetary distortions—for example, labor supply distortions and the tendency for real cash balances to be suboptimally low when the nominal interest rate is greater than zero—are quantitatively unimportant; (ii) in models where there is some motive for holding money, if we take the limit as the motive for holding money goes to zero, then this limiting economy has essentially the same properties as does the cashless economy. The purpose

of this section is to evaluate these arguments in the context of a particular monetary model.

For our purposes, a convenient expository vehicle is a cash-in-advance model of money and credit where we can parameterize the friction that makes money useful in transactions. There are other types of approaches we could take here; for example, we could use a monetary search and matching framework along the lines of Lagos and Wright (2005),[2] but the model we use here allows us to append monetary exchange to the cashless model in Section 2 with the least fuss. Our framework is much like that in Alvarez, Lucas, and Weber (2001), absent limited participation frictions, but including the labor supply decision and the sticky-price friction we have been maintaining throughout. The structure of preferences, technology, and price determination is identical to that which we assumed in the cashless model.

Here, suppose that the representative consumer trades on asset markets at the beginning of each period and then takes the remaining money to the goods market, where goods can be purchased with money and credit. The consumer faces the cash-in-advance constraint

$$P_t c_t^1 + P_{t-1} c_t^2 + q_t b_{t+1} \leq \theta W_t (n_t^1 + n_t^2) + m_t + \tau_t + s_t l_t + b_t, \qquad (43)$$

where b_t denotes one-period nominal bonds purchased by the consumer in period $t-1$, each of which pays off one unit of money in period t; m_t denotes nominal money balances carried over from the previous period; τ_t is a nominal lump-sum transfer from the government; and $s_t l_t$ is a within-period money loan from the central bank, where l_t is the nominal amount that must be returned to the central bank at the end of the period. Also, θ denotes the fraction of current-period wage income that can be accessed in the form of within-period credit when the consumer trades in the goods market. Note that $\frac{1}{s_t} - 1$ is the within-period nominal interest rate on central bank loans, and, as above, $\frac{1}{q_t} - 1$ is the one-period nominal interest rate. Here $0 \leq \theta \leq 1$, and θ is the critical parameter that captures the usefulness of money in transactions. With $\theta = 0$, this is a pure monetary economy, and with $\theta = 1$, money is irrelevant.

The consumer must also satisfy his or her budget constraint, given by

$$P_t c_t^1 + P_{t-1} c_t^2 + m_{t+1} + q_t b_{t+1} = W_t (n_t^1 + n_t^2) + m_t + \tau_t + (s_t - 1) l_t + b_t. \quad (44)$$

Let M_t denote the supply of money and L_t denote the supply of central bank loans. Then the asset market equilibrium conditions are

$$m_t = M_t; \; b_t = 0; \; l_t = L_t, \qquad (45)$$

or, money demand equals money supply, the demand for bonds equals the supply, and the demand for central bank loans equals the supply.

---

[2] See, for example, Aruoba and Schorfheide (2008), where a monetary search model with nominal rigidities is constructed for use in quantitative work.

Eliminating Intertemporal Distortions with Central Bank Lending

In general, an equilibrium in this model is difficult to characterize, as price stickiness complicates the dynamics. In part, our goal will be to determine the features of a "cashless limit" in this economy, along the lines of Woodford (2003). To that end, given the New Keynesian view that intertemporal distortions are unimportant, suppose that the regime of central bank lending is set up so that those distortions are eliminated. That is, suppose that the central bank supplies no money, except through central bank loans made at the beginning of the period at a zero nominal interest rate.

Let $s_t = 1$, and suppose that the central bank accommodates whatever demand for central bank loans arises at a zero nominal interest rate. We then have $M_t = \tau_t = 0$ for all t and $L_t = (1-\theta)W_t(n_t^1+n_t^2)$. Then, given (43)–(45), optimization, and goods market equilibrium, we can define an equilibrium in terms of relative prices and quantities, just as in the cashless economy.

This monetary regime is then one where all economic agents have access to a daylight overdraft facility, much like the Federal Reserve System uses each day to accommodate payments among financial institutions. Given a zero nominal interest rate on daylight overdrafts, money will not be held between periods, which we can interpret as a system in which holdings of outside money are zero overnight (interpreting a period as a day). This setup is extreme, as it allows universal access to central bank lending facilities and does not admit anything resembling currency-holding in equilibrium.

In equilibrium, (7)–(11) must be satisfied, just as in the cashless economy. The key difference in this monetary economy will be in the determination of π_t. Supposing for convenience that (43) always holds with equality in each period, then, given (7), we can determine π_t by

$$\pi_t = \frac{L_t w_{t-1}(n_{t-1}^1 + n_{t-1}^2)}{L_{t-1} w_t (n_t^1 + n_t^2)}, \tag{46}$$

for $t = 1, 2, 3, ...,$ with π_0 given. An equilibrium is a stochastic process $\{n_t^1, n_t^2, c_t^1, c_t^2, w_t, \pi_t, q_t\}_{t=0}^{\infty}$, with π_0 given, solving (7)–(11) and (46). The solution must satisfy $q_t \leq 1$ for all t, which assures that an equilibrium exists where (43) holds with equality. In general, it is not straightforward to characterize a solution, but it is clear from (46) that the solution is consistent with the quantity theory of money. That is, L_t is the nominal quantity of money available to spend in period t, and $w_t(n_t^1 + n_t^2) = \gamma_t n_t^1 + \frac{\gamma_t n_t^2}{\pi_t}$ is total GDP. Therefore, (46) states that the rate of increase in the price of the flexible-price good is roughly equal to the rate of money growth minus the rate of growth in real GDP. Note that the parameter θ does not appear anywhere in (7)–(11) and (46). That is, we can treat the equilibrium solution as the cashless limit, as we will obtain the same solution for any $\theta > 0$. Note here that the cashless limit of this monetary economy is *not* the cashless economy, and the quantity

of money is important for the solution, along quantity theory lines, in spite of the fact that no money is held between periods. Thus, we have followed the logic of Woodford (2003) here, but we do not get Woodford's results.

What is an optimal monetary policy here, given that the within-period nominal interest rate is zero? The key choice for the central bank is L_t, the nominal loan quantity in each period. If L_t can be set so that $\pi_t = 1$ for all t, then clearly this would be an optimal policy, since from (7)–(11) we will obtain $n_t^i = \tilde{n}_t$ for all t and $i = 1, 2$. From equation (46), this requires that

$$\frac{L_t}{L_{t-1}} = \frac{\gamma_t \tilde{n}_t}{\gamma_{t-1}\tilde{n}_{t-1}}. \tag{47}$$

This optimal policy then implies a nominal bond price

$$q_t = \beta E_t \left[\frac{u'(\gamma_{t+1}\tilde{n}_{t+1})}{u'(\gamma_t \tilde{n}_t)} \right], \tag{48}$$

and for this optimal policy to support an equilibrium, we require that $q_t \leq 1$ for all t, which is satisfied provided the variability in γ_t is sufficiently small. Thus, we can define the optimal monetary policy as a rule for monetary growth, which accommodates GDP growth according to (47) or as a nominal interest rate rule governed by (48), i.e., the money growth rule and nominal interest rate rule are flip sides of the same monetary policy. On the one hand, the money growth rule in (47) states that the money supply should always grow at the same rate as optimal GDP. On the other hand, the nominal interest rate rule states that the nominal interest rate should move in response to productivity shocks in such a way that it is equal to the optimal real interest rate.

Part of the New Keynesian justification for use of a cashless model (see Woodford 2003) is that if intertemporal frictions are insignificant, even if there is a monetary friction in the model, then the prescription for the optimal nominal interest rate rule is the same as in the cashless economy. This is certainly true here, as the nominal interest rate rule implicit in (48) is the same as (12). For our purposes, though, the monetary economy is more informative about policy, as it says something about the monetary policy regime that is necessary to get this result, and gives us a prescription for how the central bank should manipulate the quantities that it has under its control.

3. DISCUSSION

In this section we will discuss our results, organized in terms of monetary policy instruments, Phillips curves, and monetary frictions.

Monetary Policy Instruments

What can a central bank control? Ultimately, if we ignore the central bank's regulatory powers, a central bank can control two sets of things. First, it

can determine the quantities on its balance sheet, including lending to private sector economic agents. Second, it can determine the interest rates on deposits (reserves) held with the central bank and the interest rates on the loans it makes, in particular the interest rates on daylight overdrafts associated with payments made using outside money and the interest rates on overnight central bank lending or lending at longer maturities. The key power a central bank holds is its monopoly on the issue of fiat money. Essentially, a central bank is much like any other bank in that its liabilities serve as a means of payment, and it performs a type of liquidity transformation in intermediating assets that are difficult to use as means of payment. Central bankers, and many economists, hold the view that the quantity of intermediation that the central bank carries out, reflected in the quantity of fiat money outstanding, has consequences for real economic activity in the short run and for prices.

Central banks cannot set market interest rates, though they might like to. New Keynesians typically model central bank behavior as the determination of a market nominal interest rate as a function of endogenous and exogenous variables. There are good reasons to think that a central bank operating procedure consisting of periodic revision of an overnight nominal interest rate target or inflation rate targeting is preferable to the money growth targeting that Friedman (1968) had in mind. That is, the predominant shocks that are of concern to the central bank in the very short run, say between Federal Open Market Committee meetings, are financial market shocks that cause fluctuations in the demand for outside money. Given that these shocks are difficult to observe, a sensible procedure may be to smooth the overnight nominal interest rate, which may serve to optimally accommodate financial market shocks.

Though it may be possible in the short run for a central bank to use the instruments at its disposal to keep a market nominal interest rate within a tight prespecified corridor, it is inappropriate to use this as a justification for a mode of analysis that eliminates monetary considerations. A model that is used to analyze and evaluate monetary policy should tell us how the economy functions under one central bank operating procedure (e.g., monetary targeting) versus another (e.g., nominal interest rate targeting), how the instruments available to the central bank (i.e., monetary quantities) need to be manipulated to implement a particular policy rule, and how using alternative instruments (e.g., central bank lending versus open market operations) makes a difference.

Phillips Curves

Some type of Phillips curve relationship, i.e., a positive relationship between the "output gap," on the one hand, and the rate of inflation or the unanticipated component of inflation, on the other hand, is typically found in New Keynesian macroeconomic models. The Phillips curve was an important example in 1970s policy debates of how policy could go wrong in treating an empirical

correlation as structural (Lucas 1972). In New Keynesian economics, the Phillips curve has made a comeback as a structural relationship and plays a central role in reduced-form New Keynesian models (e.g., Clarida, Galí, and Gertler 1999).

As we have shown here, in a sticky-price model that seems as reasonable as typical monopolistically competitive New Keynesian setups, there is no tradeoff between output and inflation in the standard case where substitution effects dominate in the response of labor supply to a wage increase. With a zero inflation rate, output is maximized. Even in the case where income effects are large, more output can be obtained if the inflation rate deviates from zero, but this is inefficient. Further, in this case, more output can be obtained not only with inflation, but with deflation.

Monetary Frictions

We know that what makes a modern central bank unique is the power granted to it as the monopoly issuer of outside money, which takes the form of deposits with the central bank and circulating currency. We also know that central banking is not a necessity. Indeed, there are examples of economies that grew and thrived without a central bank. The United States did not have a central bank until 1914, and the private currency systems in place in Scotland from 1716–1845 and in Canada before 1935 are generally regarded as successes. Before asking how a central bank should behave, we might want to ask what justifies its existence in the first place.

From the viewpoint of a monetary economist, a theory of central banking should not only tell us what the role of the central bank is in a modern economy, but also why we should grant the central bank a monopoly in supplying key media of exchange. Such a central banking theory must necessarily come to grips with the principal frictions that make money useful as a medium of exchange and the frictions that may make private provision of some types of media of exchange inefficient.

New Keynesians argue that we can do a better job of understanding how monetary policy works and how it should be conducted by ignoring these frictions. By using a very simple cash-in-advance construct, we have shown these arguments require some very special assumptions. For our cashless sticky-price economy to work in the same way as does a comparable monetary economy requires that: (i) a monetary regime be in place that corrects intertemporal inefficiencies; (ii) all economic agents be on the receiving end of the central bank's actions; and (iii) currency holding be unimportant. While some countries, such as Canada and New Zealand, have moved to monetary systems without reserve requirements and with interest on reserves, thus correcting some distortions, most countries are far from the elimination of intertemporal monetary frictions. Thus, in practice it is likely that intertemporal

distortions play an important role, and arguably *the* important role if we are considering the effects of long-run anticipated inflation. Also, the fact that not all economic agents are on the receiving end of monetary policy actions, which gives rise to distributional effects of monetary policy, is regarded as important in the segmented markets literature. Market segmentation (in both goods and financial markets) is perhaps of greater significance than sticky-price frictions in generating short-run nonneutralities of money (see Alvarez and Atkeson 1997; Alvarez, Atkeson, and Kehoe 2003; Williamson [ForthcomingA, ForthcomingB]). Finally, currency is still widely used in the world economy (Alvarez and Lippi 2007). In spite of technological improvements in transactions technologies, currency is a wonderfully simple transactions technology that permits exchange in the many circumstances where anonymous individuals need to transact with each other.

4. CONCLUSION

Recent events involving turmoil in credit markets and heretofore unheard-of interventions by the Federal Reserve System make it abundantly clear that the monetary policy problem is far from solved. Further, for the key questions that need to be answered in the midst of this crisis, New Keynesian economics appears to be unhelpful. How is central bank lending different from open market operations in terms of the effects on financial markets and goods markets? To which institutions should a central bank be lending and under what conditions? What regulatory power should the Federal Reserve System exercise over the institutions to which it lends? Should the Fed's direct intervention be limited to conventional banks, or should this intervention be extended to investment banks and government-sponsored financial institutions? Unfortunately, typical New Keynesian models ignore credit markets, monetary frictions, and banking and are, therefore, of little or no use in addressing these pressing questions.

What hope do we have of developing a theory of money and central banking that can satisfy monetary economists and also be of practical use to central bankers? Monetary economics and banking theory have come a long way in the last 30 years or more, and perhaps the economics profession needs to be educated as to why modern monetary and banking theory is useful and can be applied to policy problems. We now understand that recordkeeping and the flow of information over space and time is critical to the role of currency as a medium of exchange (Kocherlakota 1998). We know that decentralized exchange with currency can lead to holdup problems that accentuate the welfare losses from inflation (Lagos and Wright 2005). We understand how banks act to insure private agents against liquidity risk (Diamond and Dybvig 1983) and to economize on monitoring costs (Diamond 1984, Williamson 1986). We know that financial market segmentation and goods

market segmentation are important for monetary policy (Alvarez and Atkeson 1997; Alvarez, Atkeson, and Kehoe 2003; Williamson [Forthcoming A; Forthcoming B]). Putting together elements of these ideas in a comprehensive theory of central banking is certainly within our grasp, and I very much look forward to future developments.

REFERENCES

Alvarez, Fernando, and Andrew Atkeson. 1997. "Money and Exchange Rates in the Grossman-Weiss-Rotemberg Model." *Journal of Monetary Economics* 40 (December): 619–40.

Alvarez, Fernando, Andrew Atkeson, and Patrick Kehoe. 2002. "Money, Interest Rates, and Exchange Rates with Endogenously Segmented Markets." *Journal of Political Economy* 110 (February): 73–112.

Alvarez, Fernando, and Francesco Lippi. 2007. "Financial Innovation and the Transactions Demand for Cash." University of Chicago Working Paper.

Alvarez, Fernando, Robert Lucas, and Warren Weber. 2001. "Interest Rates and Inflation." *American Economic Review Papers and Proceedings* 91: 219–25.

Aruoba, S.B., and Frank Schorfheide. 2008. "Insights from an Estimated Search-Based Monetary Model with Nominal Rigidities." University of Maryland and University of Pennsylvania Working Paper.

Calvo, Guillermo. 1983. "Staggered Prices in a Utility Maximizing Framework." *Journal of Monetary Economics* 12: 383–98.

Caplin, Andrew, and Daniel Spulber. 1987. "Menu Costs and the Neutrality of Money." *Quarterly Journal of Economics* 102: 703–26.

Clarida, Richard, Jordi Galí, and Mark Gertler. 1999. "The Science of Monetary Policy: A New Keynesian Perspective." *Journal of Economic Literature* 37 (December): 1661–1707.

Diamond, Douglas. 1984. "Financial Intermediation and Delegated Monitoring." *Review of Economic Studies* 51: 393–414.

Diamond, Douglas, and Philip Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy* 91: 401–19.

Dixit, Avinash, and Joseph Stiglitz. 1977. "Monopolistic Competition and Optimum Product Diversity." *American Economic Review* 67: 297–308.

Fischer, Stanley. 1977. "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule." *Journal of Political Economy* 85 (February): 191–205.

Friedman, Milton. 1968. "The Role of Monetary Policy." *American Economic Review* 58: 1–16.

Goodfriend, Marvin, and Robert King. 1997. "The New Neoclassical Synthesis and the Role of Monetary Policy." In *NBER Macroeconomics Annual,* Vol. 12. Chicago: University of Chicago Press, 231–83.

Hicks, John R. 1937. "Mr. Keynes and the 'Classics:' A Suggested Interpretation." *Econometrica* 5 (April): 147–59.

Kocherlakota, Narayana. 1998. "Money is Memory." *Journal of Economic Theory* 81 (August): 232–51.

Lagos, Ricardo, and Randall Wright. 2005. "A Unified Framework for Monetary Theory and Policy Analysis." *Journal of Political Economy* 113 (June): 463–84.

Lucas, Robert. 1972. "Expectations and the Neutrality of Money." *Journal of Economic Theory* 4: 103–24.

Lucas, Robert, and Nancy Stokey. 1987. "Money and Interest in a Cash-in-Advance Economy." *Econometrica* 55 (May): 491–513.

Mankiw, N. Gregory. 1985. "Small Menu Costs and Large Business Cycles: A Macroeconomic Model of Monopoly." *Quarterly Journal of Economics* 100 (May): 529–37.

Samuelson, Paul A. 1997. *Economics: The Original 1948 Edition*. New York: McGraw Hill/Irwin.

Taylor, John B. 1979. "Staggered Wage Setting in a Macro Model." *American Economic Review Papers and Proceedings* 69 (May): 108–13.

Williamson, Stephen. 1986. "Costly Monitoring, Financial Intermediation and Equilibrium Credit Rationing." *Journal of Monetary Economics* 18 (September): 159–79.

Williamson, Stephen. 2008. *Macroeconomics*. Boston: Pearson (Addison-Wesley).

Williamson, Stephen. Forthcoming A. "Monetary Policy and Distribution." *Journal of Monetary Economics*.

Williamson, Stephen. Forthcoming B. "Transactions, Credit, and Central Banking in a Model of Segmented Markets." *Review of Economic Dynamics*.

Woodford, Michael. 2003. *Interest and Prices*. Princeton, N.J.: Princeton University Press.

Economic Quarterly—Volume 94, Number 3—Summer 2008—Pages 219–233

Nominal Frictions, Relative Price Adjustment, and the Limits to Monetary Policy

Alexander L. Wolman

There are two broad classes of sticky-price models that have become popular in recent years. In the first class, prices adjust infrequently by assumption (so-called time-dependent models) and in the second class prices adjust infrequently because there is assumed to be a fixed cost of price adjustment (so-called state-dependent models). In both types of models it is common to assume that there are many goods, each produced with identical technologies. Consumers have a preference for variety, but their preferences treat all goods symmetrically. These assumptions mean that it is efficient for all goods to be produced in the same quantities. For that to happen, all goods must sell for the same price at any point in time. Assuming that price adjustment is staggered (as opposed to synchronized), the prices of all goods must be constant over time in order for all goods to be produced in the same quantities. If the aggregate price level were changing over time—even at a constant rate—then with staggered price adjustment prices would necessarily differ across goods.

If there are multiple sectors that possess changing relative technologies or that face changing relative demand conditions (because consumers' preferences are changing across goods), then in general it will no longer be efficient to produce all the goods in the same quantities. Equating marginal rates of transformation to marginal rates of substitution may require relative prices to change over time. These efficient changes in *relative* prices across sectors require *nominal* prices to change within sectors. With frictions in nominal

The views here are the author's and should not be attributed to the Federal Reserve Bank of Richmond, the Federal Reserve System, or the Board of Governors of the Federal Reserve System. For helpful comments, the author would like to thank Juan Carlos Hatchondo, Andreas Hornstein, Yash Mehra, and Anne Stilwell. E-mail: alexander.wolman@rich.frb.org.

price adjustment, nominal price changes bring with them costly misallocations within (and perhaps across) sectors.[1]

In the circumstances just described, where efficiency across sectors requires nominal price changes within a sector or sectors, a zero inflation rate for the consumption price index may no longer be the optimal prescription for monetary policy in the presence of sticky prices. Which inflation rate results in the smallest distortions from price stickiness depends on the details of the environment: chiefly, the rates of relative price change across sectors and the degree of price stickiness in each sector. To cite an extreme example, suppose there are two sectors with different average rates of productivity growth. Suppose further that the sector with low productivity growth (increasing relative price) has sticky prices, whereas the sector with high productivity growth has flexible prices. Then it would be optimal to have deflation overall. Deflation would allow the desired relative price increase to occur with zero nominal price changes for the sticky-price goods and, thus, with no misallocation from the nominal frictions.

The principles for optimal monetary policy I have discussed thus far involve only frictions associated with nominal price adjustment. In reality, monetary policy must balance other frictions as well. As the literature on the Friedman rule emphasizes, the fact that money is nearly costless to produce means that it is socially optimal for individuals to face nearly zero private costs of holding money. This requires a near-zero nominal interest rate, which corresponds to moderate deflation. Other frictions are less well understood, but may be just as important. Many central banks have mandates to achieve price stability, and the fact that my models do not necessarily support this objective does not mean it is misguided; that is, my models may still be lacking. The message of this paper is not that monetary policy should deviate from zero inflation in order to minimize distortions associated with nominal price adjustment. Rather, it is that in the presence of fundamental relative price changes and nominal price adjustment frictions, there is no monetary policy—zero inflation or otherwise—that can render those frictions costless.

Section 1 works through the optimality of price stability in a benchmark one-sector model. Section 2 describes a two-sector model where a trend in relative productivities means that all prices cannot be stabilized. In Section 2 I also display U.S. data for broad categories of consumption, which display trends in relative prices. The data show that even *on average*, it is not possible to stabilize all nominal prices; trends in relative prices mean that if the price of one category of consumption goods is stabilized then prices for the other categories must have a trend. Furthermore, relative price trends in the United

[1] These statements are qualitative ones. Unless monetary policy is extremely volatile or generates very large inflation or deflation, most current models attribute relatively small welfare costs to nominal frictions.

States have been rather large; since 1947, the price of the services component of personal consumption expenditures has risen by a factor of five relative to the price of the durable goods component. Sections 3 and 4 provide a brief review of existing literature on relative price variation and monetary policy. In contrast to the material in Section 2, the existing literature has concentrated on random fluctuations in relative prices around a steady state where relative prices are constant. Section 3 reviews the literature on cyclical variation in relative prices, and Section 4 summarizes related work on wages (a form of relative price) and prices across locations. Section 5 concludes.

1. OPTIMALITY OF PRICE STABILITY IN A ONE-SECTOR MODEL

Here I formalize the explanation for how price stability eliminates the distortions associated with price stickiness in a one-sector model. Suppose there are a large number of goods, specifically a continuum of goods indexed by $z \in [0, 1]$, and suppose that consumers' utility from consuming $c_t(z)$ of each good is given by c_t, where that utility is determined by the following aggregator function:

$$c_t = \left[\int_0^1 c_t(z)^{(\varepsilon-1)/\varepsilon} dz \right]^{\varepsilon/(\varepsilon-1)}, \tag{1}$$

where ε is the elasticity of substitution in consumption between different goods. Suppose that each good is produced with a technology that uses only labor input, and that one unit of the consumption good can be produced with $1/a_t$ units of labor input:

$$c_t(z) = a_t n_t(z), \quad \text{for } z \in [0, 1]. \tag{2}$$

Thus, a_t is a productivity factor common to all goods. I assume that a_t is exogenous. In particular, monetary policy has no effect on a_t. Finally, there is a constraint on the total quantity of labor input:[2]

$$\int_0^1 n_t(z) \leq N_t. \tag{3}$$

Without specifying anything about the structure of markets or price-setting behavior, I can discuss efficient production of consumption goods in this model. Efficiency dictates that the marginal rate of substitution in consumption be equated to the marginal rate of transformation in production. That is, the rate at which consumers trade off goods according to their preferences

[2] In models with inelastic labor supply, N_t would be a constant equal to the time endowment. Otherwise, N_t would be equal to the difference between the time endowment and the endogenous quantity of leisure.

(represented by [1]) should be equal to the rate at which the technology (represented by [2]) allows goods to be traded off against one another in production.

For the aggregator function in (1), consumers' marginal rate of substitution between any two goods, $c_t(z_0)$ and $c_t(z_1)$, is given by

$$mrs\left(c_t(z_0), c_t(z_1)\right) = \frac{\partial c_t / \partial c_t(z_0)}{\partial c_t / \partial c_t(z_1)} = \left(\frac{c_t(z_0)}{c_t(z_1)}\right)^{-1/\varepsilon}. \tag{4}$$

For the simple linear technology in (2), the marginal rate of transformation between any two goods indexed by z_0 and z_1 is unity: Reducing the labor used in the production of z_0 by one unit yields a $1/a_t$ unit reduction in $c_t(z_0)$, and transferring that labor to the production of z_1 yields an identical $1/a_t$ unit increase in $c_t(z_1)$. Given my assumptions about consumers' preferences and the technology for producing goods, equating the marginal rate of substitution to the marginal rate of transformation requires that each good, z, be produced in the same quantity. Only then can it be the case that

$$\left(c_t(z_0)/c_t(z_1)\right)^{-1/\varepsilon} = 1 \tag{5}$$

for all $z_0, z_1 \in [0, 1]$.

At this point I know that efficiency requires all goods be produced in the same quantity. Under what conditions are the allocations in sticky-price models efficient? A standard assumption in sticky-price models, and an assumption I will make here, is that each individual good is produced by a separate monopolist. Because the Dixit-Stiglitz aggregator function (1) means that each good has many close substitutes, monopoly production of each good leads to an overall market structure known as monopolistic competition.

The demand curve faced by the monopoly producer of any good, z, is

$$c_t(z) = \left(\frac{P_t(z)}{P_t}\right)^{-\varepsilon} c_t, \tag{6}$$

where P_t is the price index for the consumption basket and is given by

$$P_t = \left[\int_0^1 P_t(z)^{1-\varepsilon} dz\right]^{1/(1-\varepsilon)}. \tag{7}$$

The demand curve and the price index can be derived from the consumer's problem of choosing consumption of individual goods in order to minimize the cost of one unit of the consumption basket (see, for example, Wolman 1999).

From the demand functions and the efficiency condition, it is clear that efficiency requires all goods to have the same price. If price adjustment is infrequent (i.e., if prices are sticky) and if price adjustment is staggered across firms, then all goods can have the same price only if the aggregate price level is constant. If the price level varied over time, then changes in the price level would occur with only some firms adjusting their price, which would be inconsistent with all firms charging the same price. In somewhat

simplified form, this is the reasoning behind optimality of price stability in sticky-price models (see, for example, Goodfriend and King 1997, Rotemberg and Woodford 1997, and King and Wolman 1999).

2. TREND VARIATION IN RELATIVE PRICES

The model sketched in the previous section is a useful benchmark, but it is obviously unrealistic to suppose that the consumption goods valued by households are all "identical" in the sense of entering preferences symmetrically (1) and being produced with identical technologies (2). Departing from that benchmark, research on monetary policy in multisector sticky-price models has concentrated on the extent to which cyclical fluctuations in the determinants of relative prices interfere with the ability of monetary policy to eliminate sticky-price distortions on a period-by-period basis, and the related question of whether overall price stability remains optimal in such environments. However, more fundamental is the question of whether a trend in relative prices affects the ability of monetary policy to eliminate distortions even in steady state, and the related question of whether price stability is optimal *on average*, i.e., whether the optimal rate of inflation is zero. I consider these questions in Wolman (2008) and I draw on that analysis in what follows, emphasizing the former question (Can distortions be eliminated in a steady state?).

Theory

In contrast to the one-sector framework, suppose that consumers have Cobb-Douglas preferences over *two* composite goods, and that each of those composites has the characteristics of the single consumption aggregate (c_t) in the previous section. Here, c_t will denote the overall consumption basket comprised of the two types of goods, and $c_{1,t}$ and $c_{2,t}$ will denote the sectoral baskets each comprised of a continuum of individual goods. The overall basket is now

$$c_t = c_{1,t}^{\nu} c_{2,t}^{1-\nu}, \tag{8}$$

and the sectoral baskets are

$$c_{k,t} = \left(\left[\int_0^1 c_{k,t}(z)^{(\varepsilon-1)/\varepsilon} dz \right]^{\varepsilon/(\varepsilon-1)} \right), \quad \text{for } k = 1, 2. \tag{9}$$

As before, ε is the elasticity of substitution between individual goods within a sector. The elasticity of substitution across sectors is unity, and the sectoral expenditure shares for the two sectors are ν and $1 - \nu$. The constraint on labor input is

$$\int_0^1 n_{1,t}(z) + \int_0^1 n_{2,t}(z) \leq N_t. \tag{10}$$

Technology for producing individual goods is the same as above,

$$c_{k,t}(z) = a_{k,t} n_{k,t}(z), \quad \text{for } k = 1, 2, \tag{11}$$

except that now I allow for different levels of productivity ($a_{k,t}$) in the two sectors. Again, productivity is exogenous, or unaffected by monetary policy.

Quantities and efficiency

I can analyze efficiency just as I did in the one-sector model. However, here there are two dimensions of efficiency to be concerned with: efficiency within sectors and efficiency across sectors. Within either sector, the analysis is identical to that in the one-sector model. Efficiency within a sector requires equal production of each good,

$$\left(c_{k,t}(z_0)/c_{k,t}(z_1)\right)^{-1/\varepsilon} = 1 \text{ for } z_0, z_1 \in [0, 1], \text{ for } k = 1, 2, \tag{12}$$

because of symmetry in preferences and identical technologies. Across sectors, the marginal rate of substitution is

$$\begin{aligned} mrs\left(c_{1,t}(z_0), c_{2,t}(z_1)\right) &= \left(\frac{\nu}{1-\nu}\right)\left(\frac{c_{2,t}}{c_{1,t}}\right)\left(\frac{c_{1,t}(z_0)/c_{1,t}}{c_{2,t}(z_1)/c_{2,t}}\right)^{\frac{-1}{\varepsilon}}, \\ \text{for } z_0, z_1 &\in [0, 1], \end{aligned} \tag{13}$$

and the marginal rate of transformation is

$$mrt\left(c_{1,t}(z_0), c_{2,t}(z_1)\right) = \left(\frac{a_{2,t}}{a_{1,t}}\right), \text{ for } z_0, z_1 \in [0, 1]. \tag{14}$$

Note that in order for efficiency to hold across sectors, it must hold within sectors; if within-sector efficiency does not hold, then from (12) the marginal rate of substitution varies across combinations of z_0, z_1. With the marginal rate of transformation independent of z (from [14]), it is not possible for the marginal rate of substitution to be equated to the marginal rate of transformation for all combinations of z_0, z_1 unless there is efficiency within each sector. Efficiency within and across sectors then holds if and only if

$$c_{k,t}(z_0) = c_{k,t}(z_1) \text{ for } z_0, z_1 \in [0, 1], \text{ for } k = 1, 2, \tag{15}$$

and

$$\frac{a_{1,t}}{a_{2,t}} = \left(\frac{1-\nu}{\nu}\right)\left(\frac{c_{1,t}}{c_{2,t}}\right). \tag{16}$$

The former condition states that quantities must be identical for all goods within a sector. The latter condition states that the ratio of sectoral consumptions should be proportional to the ratio of sectoral productivities; thus, if the ratio of sectoral productivities changes over time, then the ratio of sectoral consumptions must change in order to maintain efficiency.

Prices and efficiency

As in the one-sector model, in order to determine the conditions under which efficiency holds I need to specify market structure and pricing behavior. I make analogous assumptions to the one-sector model, namely that individual goods are produced by monopolists, which implies monopolistic competition among producers.

The demand curve faced by the monopoly producer of a good z in sector k is

$$c_{k,t}(z) = \left(\frac{P_{k,t}(z)}{P_{k,t}}\right)^{-\varepsilon} c_{k,t},\ k = 1, 2, \tag{17}$$

where $P_{k,t}$ is the price index for the sector k consumption basket,

$$P_{k,t} = \left[\int_0^1 P_{k,t}(z)^{1-\varepsilon} dz\right]^{1/(1-\varepsilon)}. \tag{18}$$

The index of sector k consumption in (17) can be replaced by the appropriate demand function,

$$c_{1,t} = \nu \left(\frac{P_{1,t}}{P_t}\right)^{-1} c_t, \text{ or} \tag{19}$$

$$c_{2,t} = (1-\nu)\left(\frac{P_{2,t}}{P_t}\right)^{-1} c_t. \tag{20}$$

These demand functions, as well as the overall price index in the two-sector model (P_t), are derived from the consumer's problem of choosing sectoral consumption in order to minimize the cost of one unit of the consumption basket. The price index is given by

$$P_t = \left(\frac{P_{1,t}}{\nu}\right)^{\nu} \left(\frac{P_{2,t}}{1-\nu}\right)^{1-\nu}. \tag{21}$$

Note from (19) and (20) that the share of consumption spending (expenditure share) going to sector one (sector two) is constant and equal to ν (equal to $1-\nu$).

From the demand curves for individual goods (17) and the within-sector efficiency condition (15), it is again clear that efficiency requires all goods within a sector to have the same price:

$$P_{k,t}(z_0) = P_{k,t}(z_1) \text{ for } z_0, z_1 \in [0, 1], \text{ for } k = 1, 2. \tag{22}$$

Across sectors, because efficiency requires relative consumptions to move with relative productivities, sectoral relative prices must vary with relative productivities. Combining (16) with (19) and (20) yields

$$\frac{a_{1,t}}{a_{2,t}} = \frac{P_{2,t}}{P_{1,t}}. \tag{23}$$

Working now in terms of prices instead of quantities, conditions (22) and (23) are necessary and sufficient for efficiency.

Earlier I stated that productivity in each sector was exogenous. Now I will make the further assumption that there is a trend in the growth rate of sector one's productivity relative to sector two's productivity:

$$\frac{a_{1,t}}{a_{2,t}} = (1+\gamma)^t, \ t = 0, 1, 2, ..., \gamma > 1, \tag{24}$$

where γ is an exogenous parameter representing relative productivity growth in sector one. Substituting this relationship into the second efficiency condition (23) implies

$$\frac{P_{2,t}}{P_{1,t}} = (1+\gamma)^t; \tag{25}$$

efficiency requires the price of sector two's composite good to rise over time relative to the price of sector one's composite good. The requirement that there be a trend in the relative price $P_{2,t}/P_{1,t}$ could be satisfied with a variety of different combinations of nominal price behavior for $P_{2,t}$ and $P_{1,t}$, but each of those combinations involves at least one nominal price having a nonzero rate of change. In other words, when there is a trend in relative productivity growth across the two sectors, some nominal prices must change in order for efficiency to hold. But now there is a contradiction, because from the requirement that prices be identical for all goods within a sector (22), I can use the same reasoning as in the one-sector model of Section 1 to conclude that efficiency with price stickiness requires zero price changes within each sector. It is not possible to have both zero price changes within each sector and a nonzero rate of price change in at least one sector.

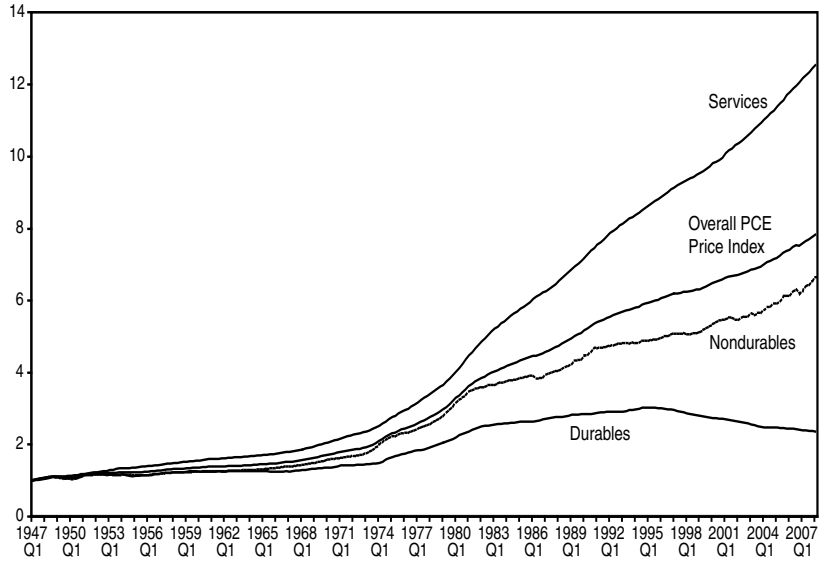
Wolman (2008) shows how one can determine the optimal rate of inflation in a sticky-price model that has the features described here. For my purpose, it is enough to conclude that when there are different trend productivity growth rates across sectors, price stickiness inevitably leads to some real distortions that cannot be undone by monetary policy.

Measurement: Price Stickiness and Relative Price Trends

It is one thing to make a theoretical argument showing that, given certain assumptions, monetary policy cannot overcome the frictions associated with price stickiness. Here I take the next step and argue that those conditions seem to exist in the United States. The conditions I discuss are, first, that there is some price stickiness (infrequent price adjustment), and second, that there are trends in the relative prices of different categories of consumption goods.

Concerning infrequent price adjustment, a vast literature has arisen in recent years reporting on the behavior of the prices of large numbers of individual

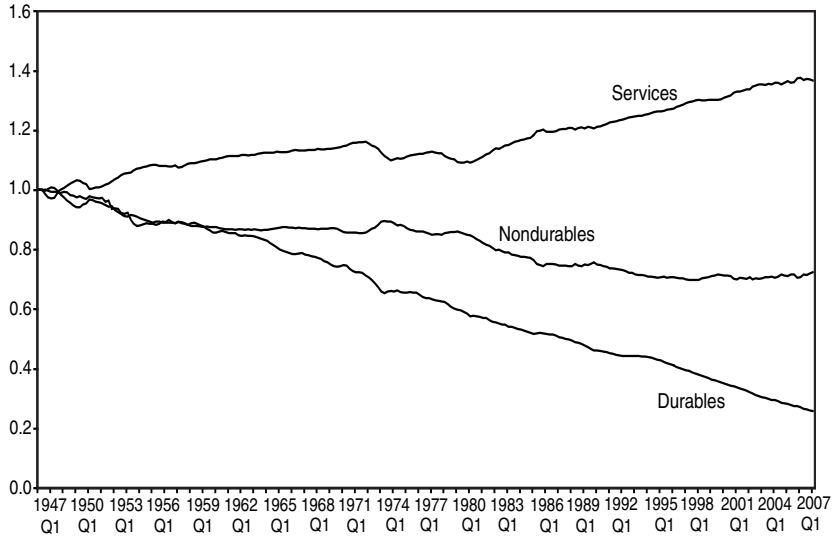
Figure 1 Sectoral Prices Indexes



goods. The seminal paper in this literature is by Bils and Klenow (2004), who study individual prices that serve as inputs into the United States Bureau of Labor Statistics (BLS) Consumer Price Index computations. Although their headline result is about price adjustment being more frequent than previous studies had estimated, they show that there is substantial heterogeneity in the frequency of price adjustment. The median price duration in their sample is 4.3 months, but one-third of consumption expenditure is on goods and services with price durations greater than 6.8 months, and one-quarter is on goods and services with price durations greater than 9 months.

In the model in Section 2, there were two sectors, and thus only one independent sectoral relative price. In contrast, the BLS compiles price indexes for hundreds of categories of consumption goods. Thus, there are hundreds of sectoral relative prices one could study. Here I will report on relative prices from highly aggregated categories of consumption: durable goods, nondurable goods, and services. Price indexes for these categories are reported by the United States Commerce Department's Bureau of Economic Analysis. Figure 1 plots the price indexes for durables, nondurables, and services from 1947 to 2008. There are clear positive trends in the prices of services and nondurables relative to durables. Since 1947, the price of nondurables relative to durables

Figure 2 "Zero-Inflation" Price Indexes



has risen by a factor of more than two, and the price of services relative to durables has risen by a factor of more than five. With these trends in relative prices, it would not have been possible to stabilize the individual prices of each consumption category. Figure 2 displays the counterfactual paths of the price indexes from Figure 1 that are consistent with zero inflation in each quarter, given the historical values for relative prices and expenditure shares; stabilizing the overall price level would have required the nominal price of durables to fall by almost 80 percent and the nominal price of services to rise by almost 40 percent.[3]

Together, the presence of infrequent individual price adjustment and trends in the relative prices of different consumption categories makes clear that monetary policy cannot eliminate the frictions associated with nominal price

[3] Figure 2 is constructed as follows: First, I normalize the levels of the sectoral price indexes at one in the first quarter of 1947. Then, for each subsequent quarter, I divide the gross rate of sectoral price change for each sector by the expenditure share weighted average of the gross rates of sectoral price changes (that is, I divide by the overall inflation rate). The resulting quotient for each sector is the rate of change of the zero-inflation price indexes in Figure 2. By construction, the expenditure share weighted average of the rates of price change for the zero-inflation sectoral price indexes is zero.

adjustment. In other words, while monetary policy can stabilize an overall price index, it cannot stabilize all individual prices. To the extent that individual price adjustment is costly, either directly (state-dependent pricing) or indirectly (time-dependent pricing), stability of individual prices is required if nominal frictions are to be eliminated.

3. CYCLICAL VARIATION IN RELATIVE PRICES

Thus far I have concentrated on the limits to monetary policy when there is a trend in the relative prices of goods with sticky nominal prices. In this setting monetary policy cannot stabilize all nominal prices, even in the absence of shocks. In contrast, most of the growing literature on relative prices and monetary policy has focused on optimal stabilization. That is, the literature has assumed that the optimal average inflation rate (the inflation target) is zero—price stability—and then gone on to study how monetary policy should make inflation behave in response to various shocks. A number of papers in this literature have used multisector models to study cyclical analogues of the kind of issues discussed in the previous section. I review some of them here.

The most influential early paper in this line of research is by Aoki (2001). In Aoki's paper there are sector-specific productivity shocks that make it infeasible to stabilize all nominal prices. However, one of the two sectors in Aoki's model has flexible prices. Thus, inability to stabilize all prices does not prevent monetary policy from neutralizing nominal frictions. Optimal monetary policy involves stabilizing prices in the sticky-price sector, allowing prices in the other sector to fluctuate with relative productivity.

Subsequently, other authors have extended Aoki's analysis to environments where price stickiness in both (or all) sectors means that monetary policy cannot neutralize the nominal frictions. Huang and Liu (2005) study a model with both intermediate and final goods sectors, with price stickiness in both sectors. As in Aoki's paper, productivity shocks are sector-specific, but they inevitably lead to distortions because of price stickiness in both sectors. Huang and Liu emphasize that stabilizing consumer prices at the expense of highly volatile producer prices can be quite costly; optimal monetary policy should place nonnegligible weight on producer price inflation.

Like Huang and Liu, Erceg and Levin (2006) study a model with two sticky-price sectors. Instead of intermediate and final goods, the sectors produce durable and nondurable final goods.[4] As in Huang and Liu's paper, sector-specific productivity shocks ought to involve relative price changes, and with price stickiness these relative price changes necessarily involve distortions. However, the presence of durable goods gives the model some

[4] Erceg and Levin's paper also includes wage stickiness, which I will discuss in the next section.

additional interesting properties. A shock to government spending—an aggregate shock—now also should involve relative price changes, because it raises the real interest rate, making durable goods less attractive. With sticky prices in both sectors, stabilizing prices in both sectors is infeasible in response to a government spending shock. Thus, even an aggregate shock can inevitably lead to nominal distortions in a multisector model.[5]

4. WAGES AND PRICES ACROSS LOCATIONS

By now it should be clear that when individual goods prices are sticky, shocks that optimally change the relative price across sectors inevitably restrict the monetary authority's ability to achieve efficient allocations. Elements of this reasoning also apply to wage stickiness and to prices across regions in a currency union.[6]

The labor market can be thought of as a sector, and if nominal prices in that sector are sticky (i.e., if nominal wages are sticky) then aggregate shocks that require real wage adjustment will lead to inefficiencies, even under optimal monetary policy. Erceg, Henderson, and Levin (2000) work through the details in a model with sticky wages and prices. Wage stickiness is introduced in a similar manner to price stickiness: Firms must assemble a range of different types of labor inputs, and the supplier of each input has monopoly power and sets her wage only occasionally. In this framework, it is not optimal to completely stabilize prices or wages, but in general higher welfare is achieved by stabilizing wage inflation than by stabilizing price inflation. Wage dispersion has two costs in the model: it makes production less efficient, and it is disliked by households, who would prefer to spread their labor input evenly over all firms. Price dispersion has only the analogue to the first cost (that is, it makes consumption of the aggregate good less efficient), and this helps to explain why wage inflation takes priority over price inflation. Another factor that works toward stabilizing wage inflation is that the productive inefficiency from wage dispersion affects each intermediate good and feeds through into inefficient production of final goods. In contrast, price dispersion leads only to inefficient production of final goods.[7]

Amano et al. (2007) use a model similar to that of Erceg, Henderson, and Levin to address trend instead of cyclical issues. They assume there is trend productivity growth so that the real wage should rise over time. In a steady state (balanced growth path), a rising real wage means that the nominal wage

[5] There are many other recent papers that study multisector models with nominal rigidities. They include Mankiw and Reis (2003, to be discussed in the next section), Carlstrom et al. (2006), Carvalho (2006), and Nakamura and Steinsson (2008).

[6] Erceg, Henderson, and Levin (2000) recognized the generality of this point.

[7] Similar reasoning may help to explain Huang and Liu's (2005) quantitative findings regarding producer price inflation, mentioned previously.

must be rising or nominal prices must be falling. With infrequent adjustment for both wages and prices, any such scenario involves distortions. Amano et al. show that optimal monetary policy involves slight deflation so that nominal wages are rising at a rate lower than real wages—a compromise between constant wages and constant prices.

Mankiw and Reis (2003) provide a general framework for thinking about optimal monetary policy in the presence of wage and price stickiness as well as sectoral considerations. They frame the monetary policy problem as choosing the appropriate index of prices and wages to stabilize. Consistent with Erceg, Henderson, and Levin (2000), Mankiw and Reis find that nominal wages carry large weight in the "price index" that should be stabilized.

Benigno (2004) studies optimal monetary policy in a two-region currency area. If nominal prices are sticky in both regions and real factors lead to efficient relative price variation across regions, then once again monetary policy cannot eliminate the real effects of price stickiness. The optimal monetary policy problem involves trading off price distortions in the two regions.

5. CONCLUSION

If prices or wages are sticky in only one sector of an economy, and if there is no heterogeneity across regions, then monetary policy can undo the effects of price stickiness. However, if there is more than one sector with sticky prices, or if wages and prices are sticky, or if there are heterogeneous regions, then nominal rigidities cause distortions under *any* monetary policy.[8] I described several examples of these distortions, emphasizing an underappreciated one, trending relative prices across sectors.

Macroeconomists are acutely aware of the limited ability of monetary policy to counteract real distortions that may be present in the economy (for example, search frictions in labor markets or monopoly power in goods markets). However, we have been perhaps less modest about the ability of monetary policy to counteract nominal distortions—in particular price adjustment frictions. A recent literature on monetary policy in multisector models with price stickiness has served to make us more modest, and this paper aims to draw attention to that literature.

[8] Loyo (2002) points out that if different sectors have different currencies, then nominal rigidities can be undone. That possibility is intriguing but not currently relevant.

REFERENCES

Amano, Robert, Kevin Moran, Stephen Murchison, and Andrew Rennison. 2007. "Trend Inflation, Wage and Price Rigidities, and Welfare." Bank of Canada Working Paper 07-42.

Aoki, Kosuke. 2001. "Optimal Monetary Policy Response to Relative Price Changes." *Journal of Monetary Economics* 48: 55–80.

Benigno, Pierpaolo. 2004. "Optimal Monetary Policy in a Currency Area." *Journal of International Economics* 63 (July): 293–320.

Bils, Mark, and Peter Klenow. 2004. "Some Evidence on the Importance of Sticky Prices." *Journal of Political Economy* 112 (October): 947–85.

Carlstrom, Charles T., Timothy S. Fuerst, Fabio Ghironi, and Kólver Hernández. 2006. "Relative Price Dynamics and the Aggregate Economy." Manuscript, http://www2.bc.edu/~ghironi/CaFuGhiroHeRelPrices082206.pdf (August).

Carvalho, Carlos. 2006. "Heterogeneity in Price Stickiness and the Real Effects of Monetary Shocks." *Frontiers of Macroeconomics,* Vol. 2, Article 1.

Erceg, Christopher, and Andrew Levin. 2006. "Optimal Monetary Policy with Durable Consumption Goods." *Journal of Monetary Economics* 53: 1341–59.

Erceg, Christopher, Dale Henderson, and Andrew Levin. 2000. "Optimal Monetary Policy with Staggered Wage and Price Contracts." *Journal of Monetary Economics* 46 (October): 281–313.

Goodfriend, Marvin, and Robert King. 1997. "The New Neoclassical Synthesis and the Role of Monetary Policy." In *NBER Macroeconomics Annual*, edited by Ben S. Bernanke and Julio Rotemberg. Cambridge, Mass.: MIT Press, 231–82.

Huang, Kevin, and Zheng Liu. 2005. "Inflation Targeting: What Inflation Rate to Target?" *Journal of Monetary Economics* 52 (November): 1435–62.

King, Robert G., and Alexander L. Wolman. 1999. "What Should the Monetary Authority do When Prices are Sticky?" In *Monetary Policy Rules,* edited by John B. Taylor. Chicago: University of Chicago Press, 349–98.

Loyo, Eduardo. 2002. "Imaginary Money Against Sticky Relative Prices." *European Economic Review* 46 (June): 1073–92.

Mankiw, N. Gregory, and Ricardo Reis. 2003. "What Measure of Inflation Should a Central Bank Target?" *Journal of the European Economic Association* 1 (September): 1058–86.

Nakamura, Emi, and Jón Steinsson. 2008. "Monetary Non-Neutrality in a Multi-Sector Menu Cost Model." Columbia University Working Paper.

Rotemberg, Julio, and Michael Woodford. 1997. "An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy." In *NBER Macroeconomics Annual*, edited by Ben S. Bernanke and Julio Rotemberg. Cambridge, Mass.: MIT Press, 297–345.

Wolman, Alexander L. 1999. "Sticky Prices, Marginal Cost, and the Behavior of Inflation." Federal Reserve Bank of Richmond *Economic Quarterly* 85 (Fall): 29–48.

Wolman, Alexander L. 2008. "The Optimal Rate of Inflation with Trending Relative Prices." Manuscript, Federal Reserve Bank of Richmond.

Economic Quarterly—Volume 94, Number 3—Summer 2008—Pages 235–263

Understanding Monetary Policy Implementation

Huberto M. Ennis and Todd Keister

Over the last two decades, central banks around the world have adopted a common approach to monetary policy that involves targeting the value of a short-term interest rate. In the United States, for example, the Federal Open Market Committee (FOMC) announces a rate that it wishes to prevail in the federal funds market, where commercial banks lend balances held at the Federal Reserve to each other overnight. Changes in this short-term interest rate eventually translate into changes in other interest rates in the economy and thereby influence the overall level of prices and of real economic activity.

Once a target interest rate is announced, the problem of implementation arises: How can a central bank ensure that the relevant market interest rate stays at or near the chosen target? The Federal Reserve has a variety of tools available to influence the behavior of the interest rate in the federal funds market (called the *fed funds rate*). In general, the Fed aims to adjust the total supply of reserve balances so that it equals demand at exactly the target rate of interest. This process necessarily involves some estimation, since the Fed does not know the exact demand for reserve balances, nor does it completely control the supply in the market.

A critical issue in the implementation process, therefore, is the sensitivity of the market interest rate to unanticipated changes in supply and/or demand.

Some of the material in this article resulted from our participation in the Federal Reserve System task force created to study paying interest on reserves. We are very grateful to the other members of this group, who patiently taught us many of the things that we discuss here. We also would like to thank Kevin Bryan, Yash Mehra, Rafael Repullo, John Walter, John Weinberg, and the participants at the 2008 Columbia Business School/New York Fed conference on "The Role of Money Markets" for useful comments on a previous draft. All remaining errors are, of course, our own. The views expressed here do not necessarily represent those of the Federal Reserve Bank of New York, the Federal Reserve Bank of Richmond, or the Federal Reserve System. Ennis is on leave from the Richmond Fed at University Carlos III of Madrid and Keister is at the Federal Reserve Bank of New York. E-mails: hennis@eco.uc3m.es, Todd.Keister@ny.frb.org.

If small estimation errors lead to large swings in the interest rate, a central bank will find it difficult to effectively implement monetary policy, that is, to consistently hit the target rate. The degree of sensitivity depends on a variety of factors related to the design of the implementation process, such as the time period over which banks are required to hold reserves and the interest rate, if any, that a central bank pays on reserve balances.

The ability to hit a target interest rate consistently plays a critical role in a central bank's communication policy. The overall effectiveness of monetary policy depends, in part, on individuals' perceptions of the central bank's actions and objectives. If the market interest rate were to deviate consistently from the central bank's announced target, individuals might question whether these deviations simply represent glitches in the implementation process or whether they instead represent an unannounced change in the stance of monetary policy. Sustained deviations of the average fed funds rate from the FOMC's target in August 2007, for example, led some media commentators to claim that the Fed had engaged in a "stealth easing," taking actions that lowered the market interest rate without announcing a change in the official target.[1] In such times, the ability to hit a target interest rate consistently allows the central bank to clearly (and credibly) communicate its policy to market participants.

Under most circumstances, the Fed changes the total supply of reserve balances available to commercial banks by exchanging government bonds or other securities for reserves in an open market operation. Occasionally, the Fed also provides reserves directly to certain banks through its discount window. In some situations, the Fed has developed other, ad hoc methods of influencing the supply and distribution of reserves in the market. For example, during the recent period of financial turmoil, the market's ability to smoothly distribute reserves across banks became partially impaired, which led to significant fluctuations in the average fed funds rate both during the day and across days. In December 2007, partly to address these problems, the Fed introduced the Term Auction Facility (TAF), a bimonthly auction of a fixed quantity of reserve balances to all banks eligible to borrow at the discount window. In principle, the TAF has increased these banks' ability to access reserves directly and, in this way, has helped ease the pressure on the market to redistribute reserves and avoid abnormal fluctuations in the market rate. Such operations, of course, need to be managed so as to achieve the ultimate goal of implementing the chosen target interest rate. Balancing the demand and supply of reserves is at the very core of this problem.

This article presents a simple analytical framework for understanding the process of monetary policy implementation and the factors that influence a

[1] See, for example, "A 'Stealth Easing' by the Fed?" (Coy 2007).

central bank's ability to keep the market interest rate close to a target level. We present this framework graphically, focusing on how various features of the implementation process affect the sensitivity of the market interest rate to unanticipated changes in supply or demand. We discuss the current approach used by the Fed, including the use of reserve maintenance periods to decrease this sensitivity. We also show how this framework can be used to study a wide range of issues related to monetary policy implementation.

In 2006, the U.S. Congress enacted legislation that will give the Fed the authority to pay interest on reserve balances beginning in October 2011.[2] We use our simple framework to illustrate how the ability to pay interest on reserves can be a useful policy tool for a central bank. In particular, we show how paying interest on reserves can decrease the sensitivity of the market interest rate to estimation errors and, thus, enable a central bank to better achieve its desired interest rate.

The model we present uses the basic approach to reserve management introduced by Poole (1968) and subsequently advanced by many others (see, for example, Dotsey 1991; Guthrie and Wright 2000; Bartolini, Bertola, and Prati 2002; and Clouse and Dow 2002). The specific details of our formalization closely follow those in Ennis and Weinberg (2007), after some additional simplifications that allow us to conduct all of our analysis graphically. Ennis and Weinberg (2007) focused on the interplay between daylight credit and the Fed's overnight treatment of bank reserves. In this article, we take a more comprehensive view of the process of monetary policy implementation and we investigate several important topics, such as the role of reserve maintenance periods, which were left unexplored by Ennis and Weinberg (2007).

1. U.S. MONETARY POLICY IMPLEMENTATION

Banks hold reserve balances in accounts at the Federal Reserve in order to satisfy reserve requirements and to be able to make interbank payments. During the day, banks can also access funds by obtaining an overdraft from their reserve accounts at the Fed. The terms by which the Fed provides daylight credit are one of the factors determining the demand for reserves by banks.

To adjust their reserve holdings, banks can borrow and lend balances in the fed funds market, which operates weekdays from 9:30 a.m. to 6:30 p.m. A bank wanting to decrease its reserve holdings, for example, can do so in this market by making unsecured, overnight loans to other banks.

The fed funds market plays a crucial role in monetary policy implementation because this is where the Federal Reserve intervenes to pursue its policy objectives. The stance of monetary policy is decided by the FOMC, which

[2] After this article was written, the effective date for the authority to pay interest on reserves was moved to October 1, 2008, by the Emergency Economic Stabilization Act of 2008.

selects a target for the overnight interest rate prevailing in this market. The Committee then instructs the Open Market Desk to adjust, via open market operations, the supply of reserve balances so as to steer the market interest rate toward the selected target.[3]

The Desk conducts open market operations largely by arranging repurchase agreements (repos) with primary securities dealers in a sealed-bid, discriminatory price auction. Repos involve using reserve balances to purchase securities with the explicit agreement that the transaction will be reversed at maturity. Repos usually have overnight maturity, but the Desk also employs other maturities (for example, two-day and two-week repos are commonly used). Open market operations are typically conducted early in the morning when the market for repos is most active.

The new reserves created in an open market operation are deposited in the participating securities dealers' bank accounts and, hence, increase the total supply of reserves in the banking system. In this way, each day the Desk tries to move the supply of reserve balances as close as possible to the level that would leave the market-clearing interest rate equal to the target rate. An essential step in this process is accurately forecasting both aggregate reserve demand and those changes in the existing supply of reserve balances that are due to *autonomous factors* beyond the Fed's control, such as payments into and out of the Treasury's account and changes in the quantity of currency in circulation. Forecasting errors will lead the actual supply of reserve balances to deviate from the intended level and, hence, will cause the market rate to diverge from the target rate, even if reserve demand is perfectly anticipated.

Reserve requirements in the United States are calculated as a proportion of the quantity of transaction deposits on a bank's balance sheet during a two-week computation period prior to the start of the maintenance period. These requirements can be met through a combination of vault cash and reserve balances held at the Fed. During the two-week reserve maintenance period, a bank's end-of-day reserve balances must, on average, equal the reserve requirement minus the quantity of vault cash held during the computation period. Reserve requirements make a large portion of the demand for reserve balances fairly predictable, which simplifies monetary policy implementation.

Reserve maintenance periods allow banks to spread out their reserve holdings over time without having to scramble for funds to meet a requirement at the end of each day. However, near the end of the maintenance period, this averaging effect tends to lose force. On the last day of the period, a bank has some level of remaining requirement that must be met on that day. This generates a fairly inelastic demand for reserve balances and makes implementing a target interest rate more challenging. For this reason, the Fed allows banks

[3] See Hilton and Hrung (2007) for a more detailed overview of the Fed's monetary policy implementation procedures.

holding excess or deficient balances at the end of a maintenance period to carry over those balances and use them to satisfy up to 4 percent of their requirement in the following period.

If a bank finds itself short of reserves at the end of the maintenance period, even after taking into account the carryover possibilities, it has several options. It can try to find a counterparty late in the day offering an acceptable interest rate. However, this may not be feasible because of an aggregate shortage of reserve balances or because of the existence of trading frictions in this market. A second alternative is to borrow at the discount window of its corresponding Federal Reserve Bank.[4] The discount window offers collateralized overnight loans of reserves to banks that have previously pledged appropriate collateral. Discount window loans are typically charged an interest rate that is 100 basis points above the target fed funds rate, although changing the size of this gap is possible and has been used, at times, as a policy instrument. Finally, if the bank does not have the appropriate collateral or chooses not to borrow at the discount window for other reasons, it will be charged a penalty fee proportional to the amount of the shortage.

Currently, banks earn no interest on the reserve balances they hold in their accounts at the Federal Reserve.[5] This situation may soon change: The Financial Services Regulatory Relief Act of 2006 allows the Fed to begin paying interest on reserve balances in October 2011. The Act also includes provisions that give the Fed more flexibility in determining reserve requirements, including the ability to eliminate the requirements altogether. Thus, this legislation opens the door to potentially substantial changes in the way the Fed implements monetary policy. To evaluate the best approach within the new, broader set of alternatives, it seems useful to develop a simple analytical framework that is able to address many of the relevant aspects of the problem. We introduce and discuss such a framework in the sections that follow.

2. THE DEMAND FOR RESERVES

In this section, we present a simple framework that is useful for understanding banks' demand for reserves. In this framework, a bank holds reserves primarily to satisfy reserve requirements, although other factors, such as the desire to make interbank payments, may also play a role. Since banks cannot fully predict the timing of payments, they face uncertainty about the net outflows from their reserve accounts and, therefore, are typically unable to exactly satisfy their reserve requirements. Instead, they must balance the possibility

[4] There are 12 regions and corresponding Reserve Banks in the Federal Reserve System. For each commercial bank, the corresponding Reserve Bank is determined by the region where the commercial bank is headquartered.

[5] See footnote 2.

of holding excess reserve balances—and the associated opportunity cost—against the possibility of being penalized for a reserve deficiency. A bank's demand for reserves results from optimally balancing these two concerns.

The Basic Framework

We assume banks are risk-neutral and maximize expected profits. At the beginning of the day, banks can borrow and lend reserves in a competitive interbank market. Let R be the quantity of reserves chosen by a bank in the interbank market. The central bank affects the supply of reserves in this market by conducting open market operations. Total reserve supply is equal to the quantity set by the central bank through its operations, adjusted by a potentially random amount to reflect unpredictable changes in autonomous factors.

During the day, each bank makes payments to and receives payments from other banks. To keep things as simple as possible, suppose that each bank will make exactly one payment and receive exactly one payment during the "middle" part of the day. Furthermore, suppose that these two payment flows are of exactly the same size, $P_D > 0$, and that this size is nonstochastic. However, the order in which these payments occur during the day is random; some banks will receive the incoming payment before making the outgoing one, while others will make the outgoing payment before receiving the incoming one.

At the end of the day, after the interbank market has closed, each bank experiences another payment shock, P, that affects its end-of-day reserve balance. The value of P can be either positive, indicating a net outflow of funds, or negative, indicating a net inflow of funds. We assume that the payment shock, P, is uniformly distributed on the interval $\left[-\overline{P}, \overline{P}\right]$. The value of this shock is not yet known when the interbank market is open; hence, a bank's demand for reserves in this market is affected by the distribution of the payment shock and not the realization.

We assume, as a starting point, that a bank must meet a given reserve requirement, K, at the end of each day.[6] If the bank finds itself holding fewer than K reserves at the end of the day, after the payment shock P has been realized, it must borrow funds at a "penalty" rate of interest, r_P, to satisfy the requirement. This rate can be thought of as the rate charged by the central bank on discount window loans, adjusted to take into account any "stigma" associated with using this facility. In reality, a bank may pay a deficiency fee instead of borrowing from the discount window or it may borrow funds

[6] We discuss more complicated systems of reserve requirements later, including multiple-day maintenance periods. For the logic in the derivations that follow, the particular value of K does not matter. The case of $K = 0$ corresponds to a system without reserve requirements.

in the interbank market very late in the day when this market is illiquid. In the model, the rate r_P simply represents the cost associated with a late-day reserve deficiency, whatever the source of that cost may be.

The specific assumptions we make about the number and size of payments that a bank sends are not important; they only serve to keep the analysis free of unnecessary complications. Two basic features of the model are important. First, the bank cannot perfectly anticipate its end-of-day reserve position. This uncertainty creates a "precautionary" demand for reserves that smoothly responds to changes in the interest rate. Second, a bank makes payments during the day as a part of its normal operations and the pattern of these payments can potentially lead to an overdraft in the bank's reserve account. We initially assume that the central bank offers daylight credit to banks to cover such overdrafts at no charge. We study the case where daylight overdrafts are costly later in this section.

The Benchmark Case

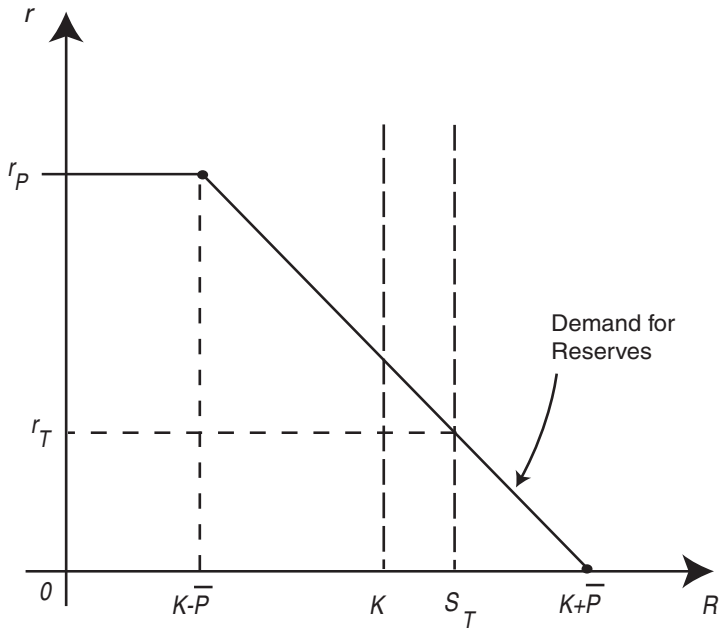
We begin by analyzing a simple benchmark case; we show later in this section how the framework can be extended to include a variety of features that are important in reality. In the benchmark case, banks must meet their reserve requirement at the end of each day, and the central bank pays no interest on reserves held by banks overnight. Furthermore, the central bank offers daylight credit free of charge.

Figure 1 depicts an individual bank's demand for reserves in the interbank market under this benchmark scenario. On the horizontal axis we measure the bank's choice of reserve holdings *before the late-day payment is realized*. On the vertical axis we measure the market interest rate for overnight loans. To draw the demand curve, we ask: Given a particular value for the interest rate, what quantity of reserves would the bank demand to hold if that rate prevailed in the market?

A bank would be unwilling to hold any reserves if the market interest rate were higher than r_P. If the market rate were higher than the penalty rate, the bank would choose to meet its requirement entirely by borrowing from the discount window. It would actually like to borrow even more than its requirement and lend the rest out at the higher market rate, but this fact is not important for the analysis. The important point is simply that there will be no demand for (nonborrowed) reserves for any interest rate larger than r_P.

When the market interest rate exactly equals the penalty rate, r_P, a bank would be indifferent between holding any amount of reserves between zero and $K - \overline{P}$ and, hence, the demand curve is horizontal at r_P. As long as the bank's reserve holdings, R, are smaller than $K - \overline{P}$, the bank will need to borrow at the discount window to satisfy its reserve requirement, K, even if the late-day inflow of funds into the bank's reserve account is the largest

Figure 1 Benchmark Demand for Reserves



possible value, $\overline{P}$.[7] The alternative would be to borrow more reserves in the market to reduce this potential need for discount window lending. Since the market rate is equal to the penalty rate, both strategies deliver the same level of profit and the bank is indifferent between them.

For market interest rates below the penalty rate, however, a bank will choose to hold at least $K - \overline{P}$ reserves. As discussed above, if the bank held fewer than $K - \overline{P}$ reserves it would be certain to need to borrow from the discount window, which would not be an optimal choice when the market rate is lower than the discount rate. The bank's demand for reserves in this situation can be described as "precautionary" in the sense that the bank chooses its reserve holdings to balance the possibility of falling short of the requirement against the possibility of ending up with extra reserves in its account at the end of the day.

[7] To see this, note that even in the best case scenario the bank will find itself holding $R+\overline{P}$ reserves after the arrival of the late-day payment flow. When $R < K - \overline{P}$, the bank's end-of-day holdings of reserves is insufficient to satisfy its reserve requirement, K, unless it takes a loan at the discount window.

If the market interest rate were very low—close to zero—the opportunity cost of holding reserves would be very small. In this case, the bank would hold enough precautionary reserves so that it is virtually certain that unforeseen movements on its balance sheet will not decrease its reserves below the required level. In other words, the bank will hold $K + \overline{P}$ reserves in this case. If the market interest rate were exactly zero, there would be no opportunity cost of holding reserves. The demand curve is, therefore, flat along the horizontal axis after $K + \overline{P}$.

In between the two extremes, $K - \overline{P}$ and $K + \overline{P}$, the demand for reserves will vary inversely with the market interest rate measured on the vertical axis; this portion of the demand curve is represented by the downward-sloping line in Figure 1. The curve is downward-sloping for two reasons. First, the market interest rate represents the opportunity cost of holding reserves overnight. When this rate is lower, finding itself with excess balances is less costly for the bank and, hence, the bank is more willing to hold precautionary balances. Second, when the market rate is lower, the relative cost of having to access the discount window is larger, which also tends to increase the bank's precautionary demand for reserves.

The linearity of the downward-sloping part of the demand curve results from the assumption that the late-day payment shock is uniformly distributed. With other probability distributions, the demand curve will be nonlinear, but its basic shape will remain unchanged. In particular, the points where the demand curve intersects the penalty rate, r_P, and the horizontal axis will be the same for any distribution with support $\left[-\overline{P}, \overline{P}\right]$.[8]

The Equilibrium Interest Rate

Suppose, for the moment, that there is a single bank in the economy. Then the demand curve in Figure 1 also represents the total demand for reserves. Let S denote the total supply of reserves in the interbank market, as jointly determined by the central bank's open market operations and autonomous factors. Then the equilibrium interest rate is determined by the height of the demand curve at point S. As shown in the diagram, there is a unique level of reserve supply, S_T, that will generate a given target interest rate, r_T.

Now suppose there are many banks in the economy, but they are all identical in that they have the same level of required reserves, face the same payment shock, etc. When there are many banks, the total demand for reserves can be found by simply "adding up" the individual demand curves. For any interest

[8] The *support* of the probability distribution is the set of values of the payment shock that is assigned positive probability. An explicit formula for the demand curve in the uniform case is derived in Ennis and Weinberg (2007). If the shock instead had an unbounded distribution, such as the normal distribution used by Whitesell (2006) and others, the demand curve would asymptote to the penalty rate and the horizontal axis but never intersect them.

rate r, total demand is simply the sum of the quantity of reserves demanded by each individual bank.

For presentation purposes, it is useful to look at the average demand for reserves, that is, the total demand divided by the number of banks. When all banks are identical, the average demand is exactly equal to the demand of each individual bank. In other words, in the benchmark case where banks are identical, the demand curve in Figure 1 also represents the *aggregate* demand for reserves, expressed in per-bank terms. The determination of the equilibrium interest rate then proceeds exactly as in the single-bank case. In particular, the market-clearing interest rate will be equal to the target rate, r_T, if and only if reserve supply (expressed in per-bank terms) is equal to S_T.

Note that the central bank has two distinct ways in which it can potentially affect the market interest rate: changing the supply of reserves available in the market and changing (either directly or indirectly) the penalty rate. Suppose, for example, that the central bank wishes to decrease the market interest rate. It could either increase the supply of reserves through open market operations, leading to a movement down the demand curve, or it could decrease the penalty rate, which would rotate the demand curve downward while leaving the supply of reserves unchanged. Both policies would cause the market interest rate to fall.

Heterogeneity

While the assumption that all banks are identical was useful for simplifying the presentation above, it is clearly a poor representation of reality in most economies. The United States, for example, has thousands of banks and other depository institutions that differ dramatically in size, range of activities, etc. We now show how the analysis above changes when there is heterogeneity among banks and, in particular, how the size distribution of banks might affect the aggregate demand for reserves.

Each bank still has a demand curve of the form depicted in Figure 1, but now these curves can be different from each other because banks may have different levels of required reserves, face different distributions of the payment shock, and/or face different penalty rates. These individual demand curves can be aggregated exactly as before: For any interest rate r, the total demand for reserves is simply the sum of the quantity of reserves demanded by each individual bank. The aggregate demand curve, expressed in per-bank terms, will again be similar to that presented in Figure 1, with the exact shape being determined by the properties of the various individual demands. If different banks have different levels of required reserves, for example, the requirement K in the aggregate demand curve will be equal to the average of the individual banks' requirements.

Our interest here is in studying how bank heterogeneity affects the properties of this demand curve. We focus on heterogeneity in bank size, which is particularly relevant in the United States, where there are some very large banks and thousands of smaller banks. We ask how large banks may differ from small banks in the context of the simple framework and how the presence of both large and small banks might affect the properties of the aggregate demand curve. To simplify the presentation, we study the three possible dimensions of heterogeneity addressed by the model one at a time. In reality, of course, the three cases are closely intertwined.

Size of Requirements

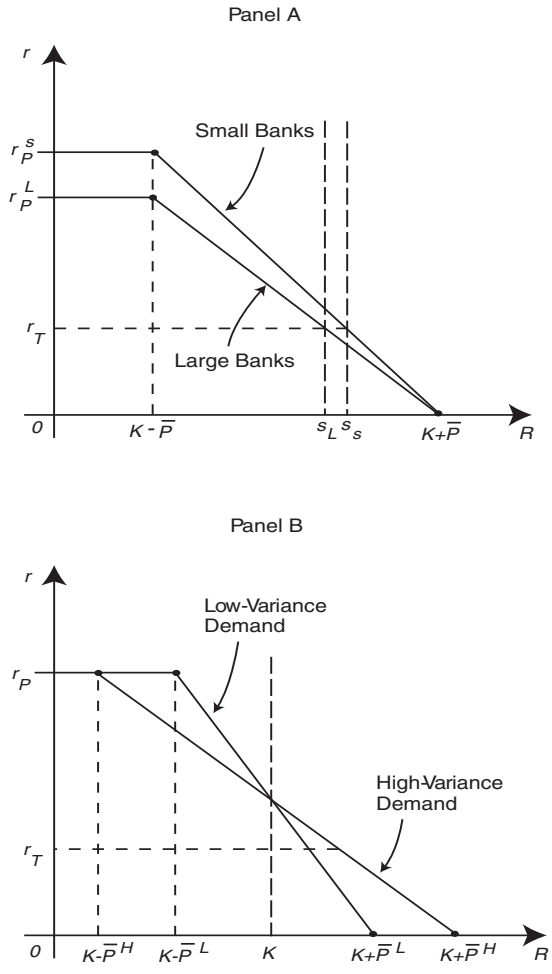
Perhaps the most natural way of capturing differences in bank size is by allowing for heterogeneity in reserve requirements. When requirements are calculated as a percentage of the deposit base, larger banks will tend to have a larger level of required reserves in absolute terms. Suppose, then, that banks have different levels of K, but they face the same late-day payment shock and the same penalty rate for a reserve deficiency. How would the size distribution of banks affect the aggregate demand for reserves in this case?

To begin, note that in Figure 1 the slope of the demand curve is independent of the size of the bank's reserve requirement, K. To see why this is the case, consider an increase in the value of K. Since both $K - \overline{P}$ and $K + \overline{P}$ become larger numbers, the demand curve in Figure 1 shifts to the right. Notice that these two points shift exactly the same distance, leaving the slope of the downward-sloping segment of the demand curve unchanged.

Simple aggregation then shows that the slope of the aggregate demand curve will be independent of the size distribution of banks. In other words, for the case of heterogeneity in K, the sensitivity of reserve demand to changes in the interest rate does not depend at all on whether the economy is comprised of only large banks or, as in the United States, has a few large banks and very many small ones.

Adding heterogeneity in reserve requirements does generate an interesting implication for the distribution of excess reserve holdings across banks. If large and small banks face similar (effective) penalty rates and are not too different in their exposure to late-day payment uncertainty, then the framework suggests that all banks should hold similar quantities of precautionary reserves. In other words, for a given level of the interest rate, the difference between the chosen reserve balances, R, and the requirement, K, should be similar for all banks. After the payment shocks are realized, of course, some banks will end up holding excess reserves and others will end up needing to borrow. On average, however, a large bank and a small one should finish the period with comparable levels of excess reserves. If the banking system is composed of a relatively small number of large banks and a much larger number of small

Figure 2 Heterogeneity



banks, then the majority of the excess reserves in the banking system will be held by small banks, simply because there are so many more of them. Even if large banks hold the majority of *total* reserve balances because of their larger requirements, most of the *excess* reserve balances will be held by small banks. This implication is broadly in line with the data for the United States.

The Penalty Rate

Another way in which small banks might differ from large ones is the penalty rate they face if they need to borrow to avoid a reserve deficiency. To be

eligible to borrow at the discount window, for example, a bank must establish an agreement with its Reserve Bank and post collateral. This fixed cost may lead some smaller banks to forgo accessing the discount window and instead borrow at a very high rate in the market (or pay the reserve deficiency fee) when necessary. Smaller banks may also have fewer established relationships with counterparties in the fed funds market and, as a consequence, may find it more difficult to borrow at a favorable interest rate late in the day (see Ashcraft, McAndrews, and Skeie 2007).

Suppose small banks do face a higher penalty rate, such as the value r_P^S depicted in Figure 2, Panel A, while larger banks face a lower rate, r_P^L. The figure is drawn as if the two banks have the same level of requirements, but this is done only to make the comparison between the curves clear. The figure shows two immediate implications of this type of heterogeneity. First, at any given interest rate, small banks will hold a higher level of precautionary reserves, that is, they will choose a larger reserve balance relative to their level of required reserves. In the figure, the smaller bank will hold a quantity S_S while the larger bank holds only S_L, even though—in this example—both face the same requirement and the same uncertainty about their end-of-day balance. As a result, the distribution of excess reserves in the economy will tend to be skewed even more heavily toward small banks than the earlier discussion would suggest.

The second implication shown in Figure 2, Panel A is that the demand curve for small banks has a steeper slope. In an economy with a large number of small banks, therefore, the aggregate demand curve will tend to be steeper, meaning that average reserve balances will be less sensitive to changes in the market interest rate. Notice that this result obtains even though there are no costs of reserve management in the model.

Support of the Payment Shock

A third way in which banks potentially differ from each other is in the distribution of the late-day payment shock they face. Figure 2, Panel B depicts two demand curves, one for a bank facing a higher variance of this distribution and one for a bank facing a lower variance. The figure shows that having more uncertainty about the end-of-day reserve position leads to a flatter demand curve and, hence, a reserve balance that is more responsive to changes in the interest rate.

In this case, it is not completely clear which curve corresponds better to large banks and which to small banks. Banks with larger and more complex operations might be expected to face much larger day-to-day variations in their payment flows. However, such banks also tend to have sophisticated reserve management systems in place. As a result, it is not clear whether the end-of-day uncertainty faced by a large bank is higher or lower than that faced

by a small bank.[9] The effect of the size distribution of banks on the shape of the aggregate demand curve is, therefore, ambiguous in this case.

Daylight Credit Fees

So far, we have proceeded under the assumption that banks are free to hold negative balances in their reserve accounts during the day and that no fees are associated with such daylight overdrafts. Most central banks, however, place some restriction on banks' access to overdrafts. In many cases, banks must post collateral at the central bank in order to be allowed to overdraft their account. The Federal Reserve currently charges an explicit fee for daylight overdrafts to compensate for credit risk. We now investigate how reserve demand changes in the basic framework when access to daylight credit is costly.

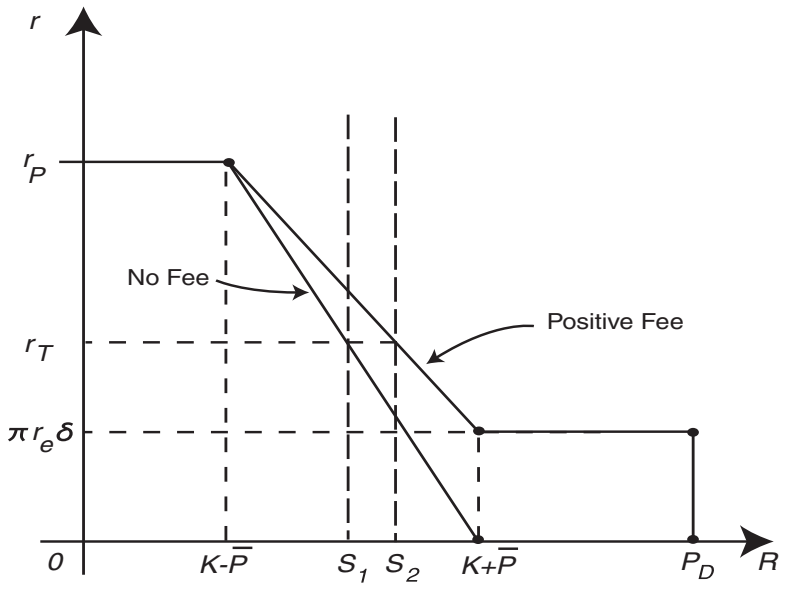
Suppose a bank sends its daytime payment, P_D, before receiving the incoming payment. If P_D is larger than R (the bank's reserve holdings), the bank's account will be overdrawn until the offsetting payment arrives. Let r_e denote the interest rate the central bank charges on daylight credit, δ denote the time period between the two payment flows during the day, and π denote the probability that a bank sends the outgoing payment before receiving the incoming one. Then the bank's expected cost of daylight credit is $\pi r_e \delta (P_D - R)$. This expression shows that an additional dollar of reserve holdings will decrease the bank's expected cost of daylight credit by $\pi r_e \delta$. In this way, the terms at which the central bank offers daylight credit can influence the bank's choice of reserve position.[10]

Figure 3 depicts a bank's demand for reserves when daylight credit is costly (that is, when $r_e > 0$). The case studied in Figure 1 (that is, when $r_e = 0$) is included in the figure for reference. It is still true that there will be no demand for reserves if the market rate is above the penalty rate r_P. The interest rate measured on the vertical axis is (as in all of our figures) the rate for a 24-hour loan. If the market rate were above the penalty rate, a bank would prefer to lend out all of its reserves at the (high) market rate and satisfy its requirements by borrowing at the penalty rate. By arranging these loans to settle at approximately the same time on both days, this plan would have no

[9] One possibility is that large banks face a wider support of the shock because of their larger operations, but face a smaller variance because of economies of scale in reserve management. This distinction cannot be captured in the figures here, which are drawn under the assumption that the distribution of the payment shock is uniform. For other distributions, the variance generally plays a more important role in the analysis than the support.

[10] The treatment of overnight reserves can, in turn, influence the level of daylight credit usage. See Ennis and Weinberg (2007) for an investigation of this effect in a closely-related framework. See, also, the discussion in Keister, Martin, and McAndrews (2008).

Figure 3 Daylight Credit Fees



effect on the bank's daylight credit usage and, hence, would generate a pure profit.

It is also still true that whenever the market rate is below the penalty rate, the bank will choose to hold at least $K - \overline{P}$ reserves, since otherwise it would be certain to need a discount window loan to meet its requirement. As the figure shows, the downward-sloping part of the demand curve is flatter when daylight credit is costly. For any market interest rate below the discount rate, the bank will choose to hold a higher quantity of reserves because these reserves now have the added benefit of reducing daylight credit fees.

Rather than decreasing all the way to the horizontal axis as in Figure 1, the demand curve now becomes flat at the bank's expected marginal cost of intraday funds, $\pi r_e \delta$. As long as R is smaller than P_D, the bank would not be willing to lend out funds at an interest rate below $\pi r_e \delta$, because the expected increase in daylight credit fees would be more than the interest earned on the loan. For values of R larger than P_D, the bank is holding sufficient reserves to cover all of its intraday payments and the demand curve drops to the horizontal axis.[11]

[11] The analysis here assumes a particular form of daylight credit usage; if an overdraft occurs, the size of the overdraft is constant over time. Alternative assumptions about the process of daytime payments would lead to minor changes in the figure, but the qualitative properties would be largely

As the figure shows, when daylight credit is costly, the level of reserves required to implement a given target rate is higher (S_2 rather than S_1 in the diagram). In other words, costly daylight credit tends to increase banks' reserve holdings. The demand curve is also flatter, meaning that reserve holdings are more sensitive to changes in the interest rate.

3. INTEREST RATE VOLATILITY

One of the key determinants of a central bank's ability to consistently achieve its target interest rate is the slope of the aggregate demand curve for reserves. In this section, we describe the relationship between this slope and the volatility of the market interest rate in the basic framework. The next two sections then discuss policy tools that can be used to limit this volatility.

While the central bank can use open market operations to affect the supply of reserves available in the market, it typically cannot completely control this supply. Payments into and out of the Treasury account, as well as changes in the amount of cash in circulation, also affect the total supply of reserves. The central bank can anticipate much of the change in such autonomous factors, but there will often be significant unanticipated changes that cause the total supply of reserves to be different from what the central bank intended. As is clear from Figure 1, if the supply of reserves ends up being different from the intended amount, S_T, the market interest rate will deviate from the target rate, r_T.

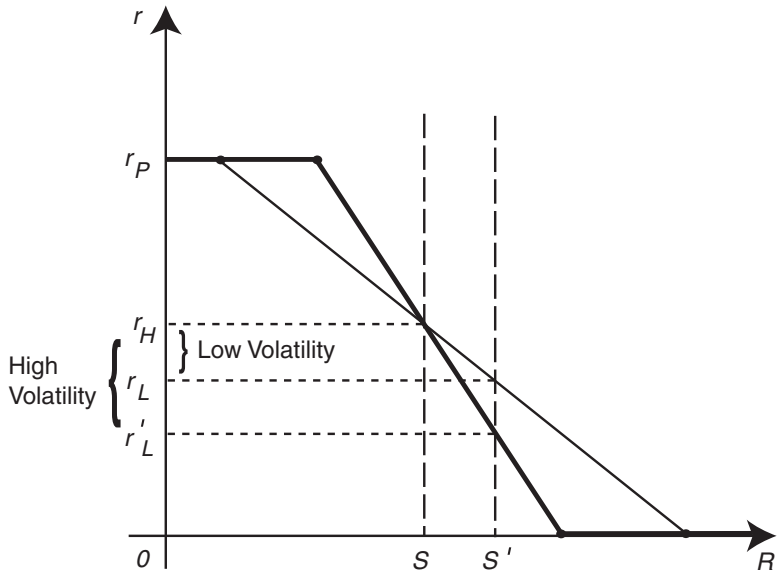
Figure 4 illustrates the fact that a flatter demand curve for reserves is associated with less volatility in the market interest rate, given a particular level of uncertainty associated with autonomous factors. Suppose this uncertainty implies that, after a given open market operation, the total supply of reserves will be equal to either S or S' in the figure. With the steeper (thick) demand curve, this uncertainty about the supply of reserves leads to a relatively wide range of uncertainty about the market rate. With the flatter (thin) demand curve, in contrast, the variation in the market rate is smaller. For this reason, the slope of the demand curve, and those policies that affect the slope, are important determinants of the observed degree of volatility of the market interest rate around the target.

As discussed in the previous section, a variety of factors affect the slope of the aggregate demand for reserves. Figure 4 can be viewed, for example, as comparing a situation where all banks face relatively little late-day uncertainty with one where all banks face more uncertainty; the latter case corresponds

---

unaffected. The analysis also takes the size and timing of payments as given. Several papers have studied the interesting question of how banks respond to incentives in choosing the timing of their outgoing payments and, hence, their daylight credit usage. See, for example, McAndrews and Rajan (2000) and Bech and Garratt (2003).

Figure 4 Interest Rate Volatility



to the thin line in the figure. However, it should be clear that the reasoning presented above does not depend on this particular interpretation. The exact same results about interest rate volatility would obtain if the demand curves had different slopes because banks face different penalty rates in the two scenarios or because of some other factor(s). What the figure shows is that there is a direct relationship between the slope of the demand curve and the amount of interest rate volatility caused by forecast errors or other unanticipated changes in the supply of reserves.

Central banks generally aim to limit the volatility of the interest rate around their target level to the extent possible. For this reason, a variety of policy arrangements have been designed in an attempt to decrease the slope of the demand curve, at least in the region that is considered "relevant." In the remainder of the article, we show how some of these tools can be analyzed in the context of our simple framework. In Section 4 we discuss reserve maintenance periods, while in Section 5 we discuss approaches that become feasible when the central bank pays interest on reserves.

4. RESERVE MAINTENANCE PERIODS

Perhaps the most significant arrangement designed to flatten the demand curve for reserves is the introduction of reserve maintenance periods. In a system

with a reserve maintenance period, banks are not required to hold a particular quantity of reserves each day. Rather, each bank is required to hold a certain *average* level of reserves over the maintenance period. In the United States, the length of the maintenance period is currently two weeks.

The presence of a reserve maintenance period gives banks some flexibility in determining when they hold reserves to meet their requirement. In general, banks will try to hold more reserves on days in which they expect the market interest rate to be lower and fewer reserves on days when they expect the rate to be higher. This flexibility implies that a bank's reserve holdings will tend to be more responsive to changes in the interest rate on any given day. In other words, having a reserve maintenance period tends to make the demand curve flatter, at least on days prior to the last day of the maintenance period. We illustrate this effect by studying a two-day maintenance period in the context of the simple framework. We then briefly explain how the same logic applies to longer periods.

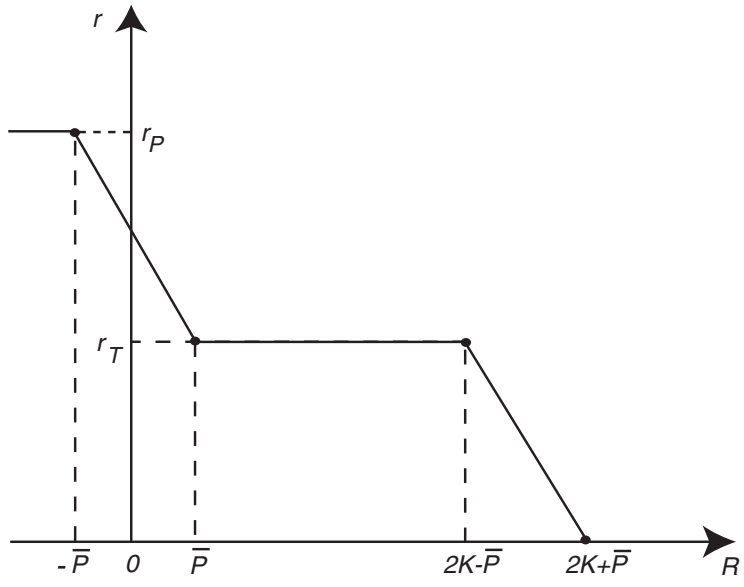
A Two-Day Maintenance Period

Let K denote the average daily requirement so that the total requirement for the two-day maintenance period is $2K$. The derivation of the demand curve for reserves on the second (and final) day of the maintenance period follows exactly the same logic as in our benchmark case. The only difference with Figure 1 is that the reserve requirement will be given by the amount of reserves that the bank has left to hold in order to satisfy the requirement for the period. In other words, the reserve requirement on the second day is equal to $2K$ minus the quantity of reserves the bank held at the end of the first day.

On the first day of the maintenance period, a bank's demand for reserves depends crucially on its belief about what the market interest rate will be on the second day. Suppose the bank expects the market interest rate on the second day to equal the target rate, r_T. Figure 5 depicts the demand for reserves on the first day under this assumption.[12] As in the basic case presented in Figure 1, there would be no demand for reserves if the market interest rate were greater than r_P. Suppose instead that the market interest rate on the first day is close to, but smaller than, the penalty rate, r_P. Then the bank will want to satisfy as much of its reserve requirement as possible on the second day, when it expects the rate to be substantially lower. However, if the bank's reserve balance after the late-day payment shock is negative, it will be forced to borrow funds at the penalty rate to avoid incurring an overnight overdraft. As long as the market rate is below the penalty rate, the bank will choose a reserve position of at least $-\overline{P}$. Note that this reserve position represents the amount of reserves held by

[12] For simplicity, Figure 5 is drawn with no discounting on the part of the bank. The effect of discounting is very small and inessential for understanding the basic logic described here.

Figure 5 A Two-Day Maintenance Period



the bank before the late-day payment shock is realized. Even if this position is negative, as would be the case when the market rate is close to r_P in Figure 5, it is still possible that the bank will receive a late-day inflow of reserves such that it does not need to borrow funds at the penalty rate to avoid an overnight overdraft. However, if the bank were to choose a position smaller than $-\overline{P}$, it would be certain to need to borrow at the penalty rate, which cannot be an optimal choice as long as the market rate is lower.

For interest rates below r_P, but still larger than the target rate, the bank will choose to hold some "precautionary" reserves to decrease the probability that it will need to borrow at the penalty rate. This precautionary motive generates the first downward-sloping part of the demand curve in the figure. As long as the day-one interest rate is above the target rate, however, the bank will not hold more than $\overline{P}$ in reserves on the first day. By holding $\overline{P}$, the bank is assured that it will have a positive reserve balance after the late-day payment shock. If the bank were holding more than $\overline{P}$ on the first day, it could lend those reserves out at the (relatively high) market rate and meet its requirement by borrowing reserves on the second day in the event that the interest rate is expected to be at the (lower) target rate, yielding a positive profit. Hence, the first downward-sloping part of the demand curve must end at $\overline{P}$.

Now suppose the first-day interest rate is exactly equal to the target rate, r_T. In this case, the bank expects the rate to be the same on both days and is,

therefore, indifferent between holding reserves on either day for the purpose of meeting reserve requirements. In choosing its first-day reserve position, the bank will consider the following issues. It will choose to hold at least enough reserves to ensure that it will not need to borrow at the penalty rate at the end of the first day. In other words, reserve holdings will be at least as large as the largest possible payment $\overline{P}$.

The bank is willing to hold more reserves than $\overline{P}$ for the purpose of satisfying some of its requirement. However, it wants to avoid the possibility of over-satisfying the requirement on the first day (that is, becoming "locked-in"), since it must hold a non-negative quantity of reserves on the second day to avoid an overnight overdraft. This implies that the bank will not be willing to hold more than the total requirement, $2K$, minus the largest possible payment inflow, $\overline{P}$, on the first day. The demand curve is flat between these two points (that is, $\overline{P}$ and $2K - \overline{P}$), indicating that the bank is indifferent between the various levels of reserves in this interval.

Finally, suppose the market interest rate on the first day is smaller than the target rate. Then the bank wants to satisfy most of the requirement on the first day, since it expects the market rate to be *higher* on the second day. In this case, the bank will hold at least $2K - \overline{P}$ reserves on the first day. If it held any less than this amount, it would be certain to have some requirement remaining on the second day, which would not be an optimal choice given that the rate will be higher on the second day. As the interest rate moves farther below the target rate, the bank will hold more reserves for the usual precautionary reasons. In this case, the bank is balancing the possibility of being locked-in after the first day against the possibility of needing to meet some of its requirement on the more-expensive second day. The larger the difference between the rates on the two days, the larger the quantity the bank will choose to hold on the first day. This trade-off generates the second downward-sloping part of the demand curve.

The intermediate flat portion of the demand curve in Figure 5 can help to reduce the volatility of the interest rate on days prior to the settlement day. As long as movements in autonomous factors are small enough such that the supply of reserves stays in this portion of the demand curve, interest rate fluctuations will be minimal. For a central bank that is interested in minimizing volatility around its target rate, this represents a substantial improvement over the situation depicted in Figure 1.[13]

[13] It should be noted that Figure 5 is drawn under the assumption that the reserve requirement is relatively large. Specifically, $K > \overline{P}$ is assumed to hold, so that the total reserve requirement for the period, $2K$, is larger than the width of the support of the late-day payment shock, $2\overline{P}$. If this inequality were reserved, the flat portion of the demand curve would not exist. In general, reserve maintenance periods are most useful as a policy tool when the underlying reserve requirements are sufficiently large relative to the end-of-day balance uncertainty.

There are, however, some issues that make implementing the target rate through reserve maintenance periods more difficult than a simple interpretation of Figure 5 might suggest. First, the position of the flat portion of the demand curve at the exact level of the target rate depends on the central bank's ability to hit the target rate (on average) on settlement day. If banks expected the settlement-day interest rate to be lower than the current target, for example, the flat portion of the first-day demand curve would also lie below the target. This issue is particularly problematic when market participants expect the central bank's target rate to change during the course of a reserve maintenance period. A second difficulty is that the flat portion of the demand curve disappears on the settlement day and the curve reverts to that in Figure 1.[14] This feature of the model indicates why market interest rates are likely to be more volatile on settlement days.

Multiple-Day Maintenance Periods

Maintenance periods with three or more days can be easily analyzed in a similar way. Consider, for example, the case of a three-day maintenance period with an average daily requirement equal to K. As before, suppose that the central bank is expected to hit the target rate on the subsequent days of the maintenance period and consider the demand for reserves on the first day. This demand will be flat between the points $\overline{P}$ and $3K - \overline{P}$. That is, the demand curve will be similar to that plotted in Figure 5, but the flat portion will be wider.

To determine the shape of the demand curve for reserves on the second day we need to know how many reserves the bank held on the first day of the maintenance period. Suppose the bank held R_1 reserves with $R_1 < 3K$. Then on the second day of the maintenance period, the demand curve for reserves would be flat between the points $\overline{P}$ and $3K - R_1 - \overline{P}$. Hence, we see that as the bank approaches the final day of the maintenance period, the flat portion of its demand curve is likely to become smaller, potentially opening the door to increases in interest rate volatility. For the interested reader, Bartolini, Bertola, and Prati (2002) provide a more thorough analysis of the implications of multiple-day maintenance periods on the behavior of the overnight market interest rate using a model similar to, but more general than, ours.

[14] In practice, central banks often use carryover provisions in an attempt to generate a small flat region in the demand curve on a settlement day. Another alternative would be to stagger the reserve maintenance periods for different groups of banks. This idea goes back to as early as the 1960s (see, for example, the discussion between Sternlight 1964 and Cox and Leach 1964 in the *Journal of Finance*). One common argument against staggering the periods is that it could make the task of predicting reserve demand more difficult. Whether the benefits of reducing settlement day variability outweigh the potential costs of staggering is difficult to determine.

5. PAYING INTEREST ON RESERVES

We now introduce the possibility that the central bank pays interest on the reserve balances held overnight by banks in their accounts at the central bank. As discussed in Section 1, most central banks currently pay interest on reserves in some form, and Congress has authorized the Federal Reserve to begin doing so in October 2011. The ability to pay interest on reserves gives a central bank an additional policy tool that can be used to help minimize the volatility of the market interest rate and steer this rate to the target level. This tool can be especially useful during periods of financial distress. For example, during the recent financial turmoil, the fed funds rate has experienced increased volatility during the day and has, in many cases, collapsed to values near zero late in the day. As we will see below, the ability to pay interest on reserves allows the central bank to effectively put a floor on the values of the interest rate that can be observed in the market. Such a floor reduces volatility and potentially increases the ability of the central bank to achieve its target rate.

In this section, we describe two approaches to monetary policy implementation that rely on paying interest on reserves: an interest rate corridor and a system with clearing bands. We explain the basic components of each approach and how each tends to flatten the demand curve for reserves.

Interest Rate Corridors

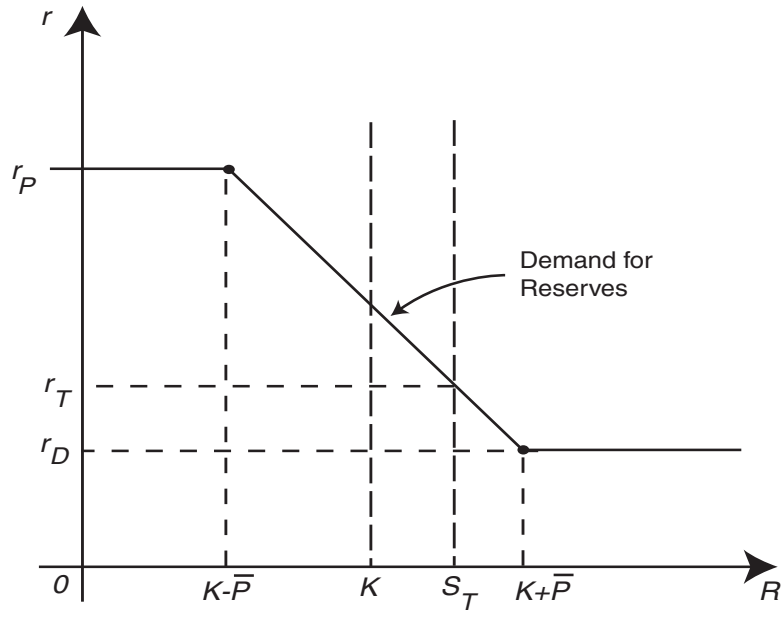
One simple policy a central bank could follow would be to pay a fixed interest rate, r_D, on all reserve balances that a bank holds in its account at the central bank.[15] This policy places a floor on the market interest rate: No bank would be willing to lend reserves at an interest rate lower than r_D, since they could instead earn r_D by simply holding the reserves on deposit at the central bank. Together, the penalty rate, r_P, and the deposit rate, r_D, form a "corridor" in which the market interest rate will remain.[16]

Figure 6 depicts the demand for reserves under a corridor system. As in the earlier figures, there is no demand for reserves if the market interest rate is higher than the penalty rate, r_P. For values of the market interest rate below r_P, a bank will choose to hold at least $K - \overline{P}$ reserves for exactly the same

[15] In practice, reserve balances held to meet requirements are often compensated at a different rate than those that are held in excess of a bank's requirement. For the daily process of targeting the overnight market interest rate, the rate paid on *excess* reserves is what matters; this is the rate we denote r_D in our analysis.

[16] A central bank may prefer to use a *lending facility* that is distinct from its discount window to form the upper bound of the corridor. Banks may be reluctant to borrow from the discount window, which serves as a lender of last resort, because they fear that others would interpret this borrowing as a sign of poor financial health. The terms associated with the lending facility could be designed to minimize this type of stigma effect and, thus, create a more reliable upper bound on the market interest rate.

Figure 6 A Conventional Corridor



reason as in Figure 1: if it held a lower level of reserves, it would be certain to need to borrow at the penalty rate, r_P. Also as before, the demand for reserves is downward-sloping in this region. The big change from Figure 1 is that the demand curve now becomes flat at the deposit rate. If the market rate were lower than the deposit rate, a bank's demand for reserves would be essentially infinite, as it would try to borrow at the market rate and earn a profit by simply holding the reserves overnight.

The figure shows that, regardless of the level of reserve supply, S, the market interest rate will always stay in the corridor formed by the rates r_P and r_D. The width of the corridor, $r_P - r_D$, is then a policy choice. Choosing a relatively narrow corridor will clearly limit the range and volatility of the market interest rate. Note that narrowing the corridor also implies that the downward-sloping part of the demand curve becomes flatter (to see this, notice that the boundary points $K - \overline{P}$ and $K + \overline{P}$ do not depend on r_P or r_D). Hence, the size of the interest rate movement associated with a given shock to an autonomous factor is smaller, even when the shock is small enough to keep the rate within the corridor.

An interesting case to consider is one in which the lending and deposit rates are set the same distance on either side of the target rate (x basis points above and below the target, respectively). This system is called *a symmetric*

corridor. A change in policy stance that involves increasing the target rate, then, effectively amounts to changing the levels of the lending and deposit rates, which shifts the demand curve along with them. The supply of reserves needed to maintain a higher target rate, for example, may not be lower. In fact— perhaps surprisingly—in the simple model studied here, the target level of the supply of reserves would not change at all when the policy rate changes.

If the demand curve in Figure 6 is too steep to allow the central bank to effectively achieve its goal of keeping the market rate close to the target, a corridor system could be combined with a reserve maintenance period of the type described in Section 4. The presence of a reserve maintenance period would generate a flat region in the demand curve as in Figure 5. The features of the corridor would make the two downward-sloping parts of the demand curve in Figure 5 less steep, which would limit the interest rate volatility associated with events where reserve supply exits the flat region of the demand curve, as well as on the last day of the maintenance period when the flat region is not present.

Another way to limit interest rate volatility is for the central bank to set the deposit rate equal to the target rate and then provide enough reserves to make the supply, S_T, intersect the demand curve well into the flat portion of the demand curve at rate r_D. This "floor system" has been recently advocated as a way to simplify monetary policy implementation (see, for example, Woodford 2000, Goodfriend 2002, and Lacker 2006). Note that such a system does not rely on a reserve maintenance period to generate the flat region of the demand curve, nor does it rely on reserve requirements to induce banks to hold reserves. To the extent that reserve requirements, and the associated reporting procedures, place significant administrative burdens on both banks and the central bank, setting the floor of the corridor at the target rate and simplifying, or even eliminating, reserve requirements could potentially be an attractive system for monetary policy implementation.

It should be noted, however, that the market interest rate will always remain some distance above the floor in such a system, since lenders in the market must be compensated for transactions costs and for assuming some counterparty credit risk. In other words, in a floor system the central bank is able to fully control the risk-free interest rate, but not necessarily the market rate. In normal times, the gap between the market rate and the rate paid on reserves would likely be stable and small. In periods of financial distress, however, elevated credit risk premia may drive the average market interest rate significantly above the interest rate paid on reserves. Our simple model abstracts from these important considerations.[17]

[17] The central bank could also set an upper limit for the quantity of reserves on which it would pay the target rate of interest to a bank; reserves above this limit would earn a lower rate (possibly zero). Whitesell (2006) proposed that banks be allowed to choose their own upper

Clearing Bands

Another approach to generating a flat region in the demand curve for reserves is the use of daily clearing bands. This approach does not rely on a reserve maintenance period. Instead, the central bank pays interest on a bank's reserve holdings at the target rate, r_T, as long as those holdings fall within a pre-specified band. Let $\underline{K}$ and $\overline{K}$ denote the lower and upper bounds of this band, respectively. If the bank's reserve balance falls below $\underline{K}$, it must borrow at the penalty rate, r_P, to bring its balance up to at least $\underline{K}$. If, on the other hand, the bank's reserve balance is higher than $\overline{K}$, it will earn the target rate, r_T, on all balances up to $\overline{K}$ but will earn a lower rate, r_E, beyond that bound.

The demand curve for reserves under such a system is depicted in Figure 7. The figure is drawn under the assumption that the clearing band is fairly wide relative to the support of the late-day payment shock. In particular, we assume that $\underline{K} + \overline{P} < \overline{K} - \overline{P}$. Let us call the interval $\left[\underline{K} + \overline{P}, \overline{K} - \overline{P}\right]$ the "intermediate region" for reserves. By choosing any level of reserves in this intermediate region, a bank can ensure that its end-of-day reserve balance will fall within the clearing band. The bank would then be sure that it will earn the target rate of interest on all of the reserves it ends up holding overnight.

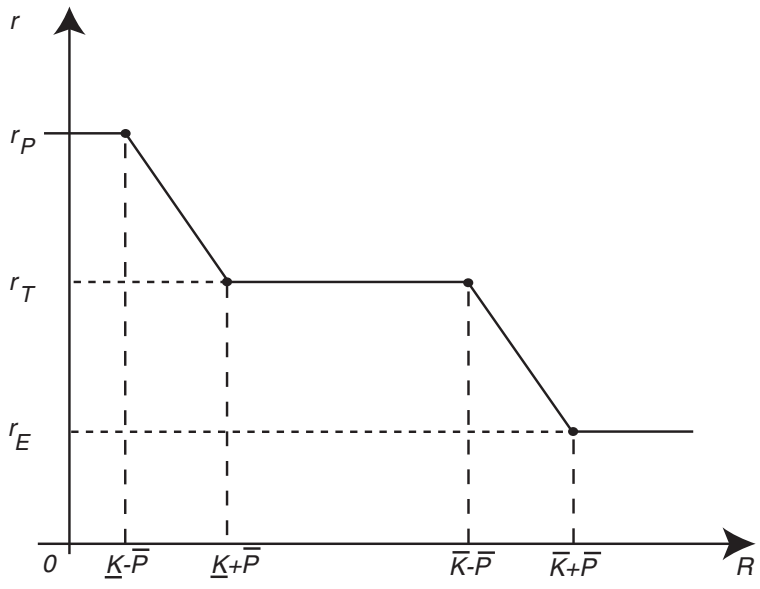
When the market interest rate is equal to the target rate, r_T, a bank is indifferent between choosing any level of reserves in the intermediate region. For example, if the bank borrows in the market to slightly increase its reserve holdings, the cost it would pay in the market for those reserves would be exactly offset by the extra interest it would earn from the central bank. Similarly, lending out reserves to slightly decrease the bank's holdings would also leave profit unchanged. This reasoning shows that the demand curve for reserves will be flat in the intermediate region between $\underline{K} + \overline{P}$ and $\overline{K} - \overline{P}$. As long as the central bank is able to keep the supply of reserves within this region, the market interest rate will equal the target rate, r_T, regardless of the exact level of reserve supply.

Outside the intermediate region, the logic behind the shape of the demand curve is very similar to that explained in our benchmark case. When the market interest rate is higher than r_T, a bank can earn more by lending reserves in the market than by holding them on deposit at the central bank. It would, therefore, prefer not to hold more than the minimum level of reserves needed to avoid being penalized, $\underline{K}$. Of course, the bank would be willing to hold some precautionary reserves to guard against the possibility that the late-day payment shock will drive their reserve balance below $\underline{K}$. The quantity of precautionary reserves it would choose to hold is, as before, an inverse function of the market interest rate; this reasoning generates the first downward-sloping part of the demand curve in Figure 7.

---

limits by paying a facility fee per unit of capacity. Such an approach leads to a demand curve for reserves that is flat at the target rate over a wide region.

Figure 7 A Clearing Band



When the market rate is below r_T, on the other hand, the bank would like to take full advantage of its ability to earn the target interest rate by holding reserves at the central bank. It would, however, take into consideration the possibility that a late-day inflow of funds will leave it with a final balance higher than $\overline{K}$, in which case it would earn the lower interest rate, r_E, on the excess funds. The resulting decision process generates a downward-sloping region of the demand curve between the rates r_T and r_E. As in Figure 6, the demand curve never falls below the interest rate paid on excess reserves (now labeled r_E); thus, this rate creates a floor for the market interest rate.

The demand curve in Figure 7 has the same basic shape as the one generated by a reserve maintenance period, which was depicted in Figure 5. It is important to keep in mind, however, that the forces generating the flat portion of the demand curve in the intermediate region are fundamentally different in the two cases. The reserve maintenance period approach relies on intertemporal arbitrage: banks will want to hold more reserves on days when the market interest rate is low and fewer reserves when the market rate is high. This activity will tend to equate the current market interest rate to the expected future rate (as long as the supply of reserves is in the intermediate region). The clearing band system relies instead on intraday arbitrage to generate the flat portion of the demand curve: banks will want to hold more reserves when

the market interest rate is low, for example, simply to earn the higher interest rate paid by the central bank.

The intertemporal aspect of reserve maintenance periods has two clear drawbacks. First, if—for whatever reason—the expected future rate differs from the target rate, r_T, it becomes difficult for the central bank to achieve the target rate in the current period. Second, large shocks to the supply of reserves on one day can have spillover effects on subsequent days in the maintenance period. If, for example, the supply of reserves is unusually high one day, banks will satisfy an unusually large portion of their reserve requirements and, as a result, the flat portion of the demand curve will be smaller on all subsequent days, increasing the potential for rate volatility on those days.

The clearing band approach, in contrast, generates a flat portion in the demand curve that always lies at the current target interest rate, even if market participants expect the target rate to change in the near future. Moreover, the width of the flat portion is "reset" every day; it does not depend on past events. These features are important potential advantages of the clearing band approach. We should again point out, however, that our simple model has abstracted from transaction costs and credit risk. As with the floor system discussed above, these considerations could result in the average market interest rate being higher than the rate r_T, as the latter represents a risk-free rate.

6. CONCLUSION

A recent change in legislation that allows the Federal Reserve to pay interest on reserves has renewed interest in the debate over the most effective way to implement monetary policy. In this article, we have provided a basic framework that can be useful for analyzing the main properties of the various alternatives. While we have conducted all our analysis graphically, our simplifying assumptions permit a fairly precise description of the alternatives and their effectiveness at implementing a target interest rate.

Many extensions of our basic framework are possible and we have analyzed several of them in this article. However, some important issues remain unexplored. For example, we only briefly mentioned the difficulties that fluctuations in aggregate credit risk can introduce in the implementation process. Also, as the debate continues, new questions will arise. We believe that the framework introduced in this article can be a useful first step in the search for much-needed answers.

REFERENCES

Ashcraft, Adam, James McAndrews, and David Skeie. 2007. "Precautionary Reserves and the Interbank Market." Mimeo, Federal Reserve Bank of New York (July).

Bartolini, Leonardo, Giuseppe Bertola, and Alessandro Prati. 2002. "Day-To-Day Monetary Policy and the Volatility of the Federal Funds Interest Rate." *Journal of Money, Credit, and Banking* 34 (February): 137–59.

Bech, M. L., and Rod Garratt. 2003. "The Intraday Liquidity Management Game." *Journal of Economic Theory* 109 (April): 198–219.

Clouse, James A., and James P. Dow, Jr. 2002. "A Computational Model of Banks' Optimal Reserve Management Policy." *Journal of Economic Dynamics and Control* 26 (September): 1787–814.

Cox, Albert H., Jr., and Ralph F. Leach. 1964. "Open Market Operations and Reserve Settlement Periods: A Proposed Experiment." *Journal of Finance* 19 (September): 534–9.

Coy, Peter. 2007. "A 'Stealth Easing' by the Fed?" *BusinessWeek*. http://www.businessweek.com/investor/content/aug2007/pi20070817_445336.htm [17 August].

Dotsey, Michael. 1991. "Monetary Policy and Operating Procedures in New Zealand." Federal Reserve Bank of Richmond *Economic Review* (September/October): 13–9.

Ennis, Huberto M., and John A. Weinberg. 2007. "Interest on Reserves and Daylight Credit." Federal Reserve Bank of Richmond *Economic Quarterly* 93 (Spring): 111–42.

Goodfriend, Marvin. 2002. "Interest on Reserves and Monetary Policy." Federal Reserve Bank of New York *Economic Policy Review* 8 (May): 77–84.

Guthrie, Graeme, and Julian Wright. 2000. "Open Mouth Operations." *Journal of Monetary Economics* 46 (October): 489–516.

Hilton, Spence, and Warren B. Hrung. 2007. "Reserve Levels and Intraday Federal Funds Rate Behavior." Federal Reserve Bank of New York Staff Report 284 (May).

Keister, Todd, Antoine Martin, and James McAndrews. 2008. "Divorcing Money from Monetary Policy." Federal Reserve Bank of New York *Economic Policy Review* 14 (September): 41–56.

Lacker, Jeffrey M. 2006. "Central Bank Credit in the Theory of Money and Payments." Speech. http://www.richmondfed.org/news_and_speeches/presidents_speeches/index.cfm/2006/id=88/pdf=true.

McAndrews, James, and Samira Rajan. 2000. "The Timing and Funding of Fedwire Funds Transfers." Federal Reserve Bank of New York *Economic Policy Review* 6 (July): 17–32.

Poole, William. 1968. "Commercial Bank Reserve Management in a Stochastic Model: Implications for Monetary Policy." *Journal of Finance* 23 (December): 769–91.

Sternlight, Peter D. 1964. "Reserve Settlement Periods of Member Banks: Comment." *Journal of Finance* 19 (March): 94–8.

Whitesell, William. 2006. "Interest Rate Corridors and Reserves." *Journal of Monetary Economics* 53 (September): 1177–95.

Woodford, Michael. 2000. "Monetary Policy in a World Without Money." *International Finance* 3 (July): 229–60.

Economic Quarterly—Volume 94, Number 3—Summer 2008—Pages 265–300

CEO Compensation: Trends, Market Changes, and Regulation

Arantxa Jarque

Compensation figures for the top managers of large firms are on the news frequently. Newspapers report the salaries, the bonuses, and the profits from selling stock options of the highest paid executives, often under headlines suggesting excessive levels of pay or a very weak relation of pay to the performance of the firms.[1] Especially after the fraud scandals at Enron and other important corporations around the world, executive performance and pay have been carefully scrutinized by the public.

Academic economists, however, have long ago recognized the importance of understanding the issues involved in determining executive pay and have been studying them for decades. In short, the main economic problem behind executive compensation design is that firm owners need to align the incentives of the executives to their own interests—typically to maximize firm value. To achieve this alignment, the compensation of the manager is usually made contingent on the performance of the firm. While in the largest firms executive compensation typically represents a small fraction of the total firm value, the decisions that a top executive makes can be potentially important for firm performance. Therefore, the way in which the dependence of compensation on firm performance is structured can have a significant impact on the added value that the executive brings to the firm. There is by now a large body of academic studies that document practices in executive pay and study the

The author would like to thank Brian Gaines, Borys Grochulski, Leornardo Martinez, and John Weinberg for helpful comments. Jarque is an assistant professor at the Universidad Carlos III de Madrid. This article was written while the author was a visiting economist at the Federal Reserve Bank of Richmond. The views expressed in this article do not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: ajarque@eco.uc3m.es.

[1] Some recent examples, following the release of an Associated Press study for 2007, include Beck and Fordahl (2008a, 2008b) and Associated Press (2008).

optimal design of compensation contracts. Some of the most important and recent findings in the literature are summarized in this article.

In Section 1, I present the various instruments commonly used to compensate executives and the main empirical regularities about executive pay. Over the last two decades, the average pay of a chief executive officer (CEO) working in one of the 500 largest firms in the United States has increased six-fold. This increase has occurred simultaneously with a change in the composition of pay of these CEOs, moving away from salary and increasingly toward performance-based compensation in the form of stock grants and stock option grants. This shift has resulted in a clear increase in the sensitivity of the total pay of CEOs to the performance of their firms. Although the increase in the level and the sensitivity are most likely related, I separate the discussion into two sections for a more detailed exposition of the evidence and a discussion of possible explanations.

In Section 2, I summarize how the level of pay has evolved since 1936. Then I briefly review some of the explanations in the academic literature for the sharp increase in the last two decades. I focus on a recent strand of the literature that has built on ideas from the classic papers of Lucas (1978) and Rosen (1981, 1982) that model firms' competition for the scarce talent of managers. These studies argue that the sharp increase in firm size and value of the last decades could be important to explain the six-fold increase in the level of CEO pay over the same period.

In Section 3, I focus on how the pay of CEOs depends on their performance. I introduce several measures of sensitivity of the CEO's pay to the results of the firm that are widely used in the literature. The empirical studies provide a wide range of estimates for sensitivity, but there is consensus about two facts: an increase in sensitivity over the last two decades and a negative relation of sensitivity with firm size (as measured by market capitalization value). I discuss some of the recent explanations for these regularities. Many of the explanations are based on studying the interaction between the manager and the owners of the firm as a moral hazard problem. In the model, shareholders minimize the cost of bringing the (risk averse) CEO to exert an unobservable effort that improves the results of the firm. This analysis is typically performed in partial equilibrium and aims to describe the form of optimal compensation contracts, i.e., the optimal sensitivity of pay to firm performance. Some recent studies suggest that a seemingly low sensitivity of pay to performance, as well as the negative relation between sensitivity of pay and firm size, could be features of dynamic optimal contracts that are used to solve the moral hazard problem. Also in this moral hazard framework, several recent articles demonstrate the efficiency of some seemingly unintuitive pay practices that are often discussed by the media, such as bonuses in bad earning years or repricing of out-of-the-money options. The level of market competition across firms, as well as the relative demand of general

versus firm-specific skills, has also been shown to be empirically significant in explaining pay trends.

In Section 4, I describe the main regulatory changes affecting executive compensation in the last 15 years: changes in personal, corporate, and capital gains taxes; new limits on the deductibility of CEO pay expenses that favor performance-based compensation; an increase in the disclosure requirements about CEO pay; a standardization of the expensing of option grants; and several initiatives fostering shareholders' activism and independence of the board of directors. Amidst the headlines on excessive pay, the popular press has been debating these changes in regulation, their potential role in the recent rise in pay, and the need for new government intervention.[2]

To shed some light on the role of regulation, I review the findings of academic studies that have rigorously tried to quantify the effect of several specific measures on the level and sensitivity of compensation. As it turns out, these studies find little evidence of a sizable effect on pay practices coming from tax advantages or salary caps. Following the main regulatory changes, some studies find evidence of a small shift of compensation from salaries and bonuses toward performance-based compensation (stocks and options), which translates into a slight increase in the sensitivity of pay to performance. Better corporate governance and the increase in the proportion of institutional shareholders appear to be associated with higher company returns and higher incentives for CEOs. This suggests that regulation efforts to improve corporate governance and transparency have been moving in the right direction, although it is difficult to evaluate the relative importance of regulation versus the market-induced changes in governance practices.

Designing executive compensation packages is, no doubt, complicated. Judging the appropriateness of those packages is, consequently, a very difficult task in which models and sophisticated econometric tools are a necessity. I now proceed to review the most recent attempts at this task and their conclusions.

1. UNDERSTANDING AND MEASURING CEO COMPENSATION

Today companies pay their top executives through some or all of the following instruments: a salary, a bonus program, stock grants (usually with restrictions

[2] On one hand, in the most recent special report on CEO compensation published by the *Wall Street Journal* (2008), one of the articles is dedicated to the unintended consequences of regulation changes. On the other hand, the "say on pay" proposals, which seek to force boards of directors to have a (nonbinding) vote on the compensation plan that they design for a CEO, have been receiving particular attention in the press—see *The Economist* (2008b, "Pay Attention") on the recent trend of pay in Europe and the regulation responses of some European governments, and *The Economist* (2008a, "Fair or Foul") on the "say on pay" proposals both in Europe and the United States, which cites both U.S. presidential candidates as being in favor of forcing the shareholders' vote.

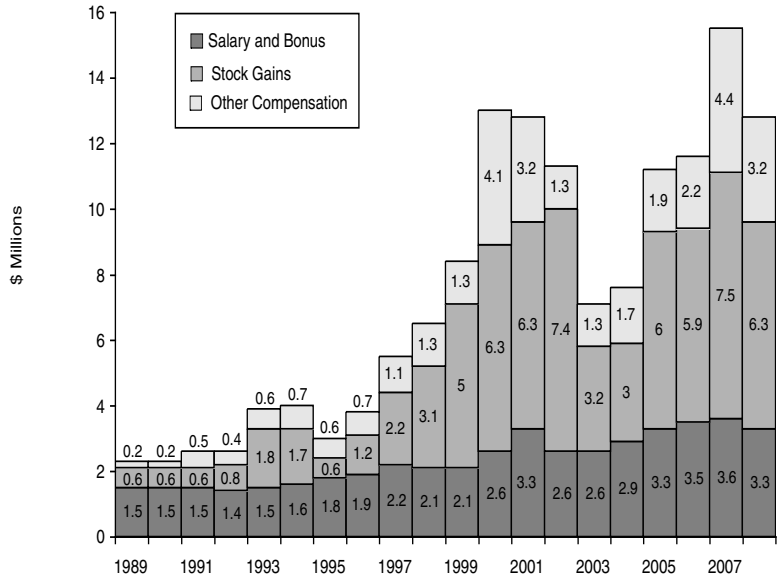
on the ability to sell them), grants of options on the stock of the firm, and long-term incentive plans that specify retirement and severance payments, as well as pension plans and deferred benefits. The most accepted explanation for these variable compensation instruments is the existence of a moral hazard problem: The separation of ownership and control of the firm implies the need to provide incentives to the CEO that align his interests with those of the firm owners. In the presence of moral hazard, the optimal contract prescribes that the pay of the executive should vary with the results of the firm. However, in spite of the need for incentives, limited funds on the part of CEOs or risk aversion considerations imply that exposing the CEO to the same risk as shareholders is typically either an unfeasible or inefficient arrangement. The optimal contract should balance incentives and insurance. Therefore, part of the compensation should be variable and could be provided through stock or stock options, while part of the compensation, such as the annual salary, should not be subject to risk, providing some insurance to the CEO against bad firm performance due to factors that he cannot control.

Data on the level of annual compensation of workers classified as CEOs are available from the Bureau of Labor Statistics (BLS). These are wages representative of the whole economy. However, details are not available on the specific forms of the contract (i.e., the compensation instruments that were used in providing that compensation). Data from the BLS show that the average CEO earns an annual wage of $151,370 (May 2007)—less than the average doctor (internist) and about $25,000 more per year than a lawyer. The annual wage of the average CEO today is about 3.5 times that of the average worker in the economy, and the evolution of this comparison over the last seven years has followed a similar pattern as that of other white-collar professions. However, the distribution of wages of CEOs is extremely skewed. Figure 1 shows the evolution of the total level of pay for the CEOs of the 500 largest firms in the economy, according to the figures in *Forbes*, a magazine that publishes surveys of top CEO pay annually. The average annual wage of this sample of executives in 2007 was about $15 million, approximately 300 times that of the average worker in the U.S. economy.

For the top 500 firms, information about CEO pay is readily available. Since these companies are public, they file their proxy statements with the Securities and Exchange Commission (SEC), making their reports on their top executives' compensation public. Most academic studies restrict themselves to studying the compensation of the CEOs in this subsample of firms. Therefore, I too concentrate on this subsample in the rest of the article. From this point on, the acronym "CEO" refers to the chief executive officer of one of the top 500 firms in the United States.

Figure 1 also shows the evolution of the three main components of CEO pay from 1989 to 2007. Stock gains refers to the value realized during the given fiscal year by exercising vested options granted in previous years. The

Figure 1 Compensation Data Based on *Forbes* Magazine's Annual Survey (2007 Constant Dollars)



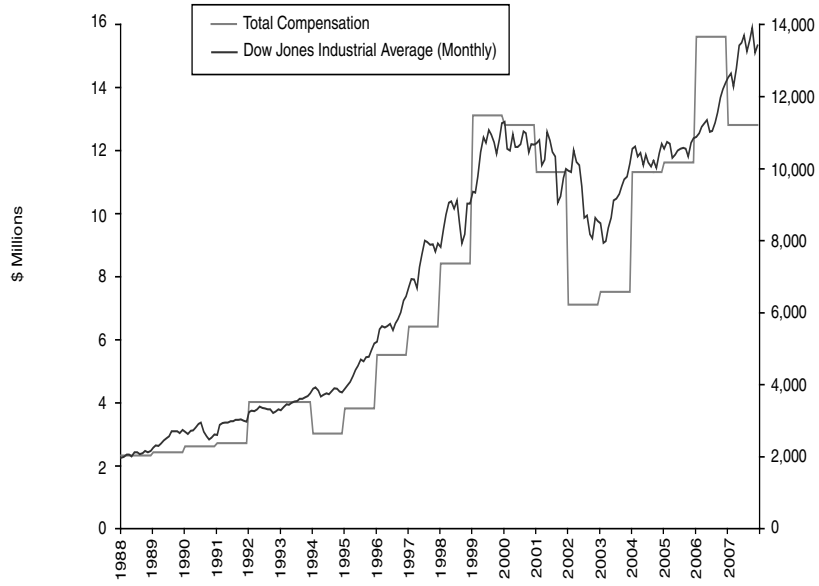
gain is the difference between the stock price on the date of exercise and the exercise price of the option. Other compensation includes long-term incentive payouts and the value realized from vesting of restricted stock and performance shares, and other perks such as premiums for supplemental life insurance, annual medical examinations, tax preparation and financial counseling fees, club memberships, security services, and the use of corporate aircraft. The main trend that we observe in this figure is that, although annual salaries have been increasing, the proportion of total pay that they represent has decreased in the last 20 years, while compensation through options has become the most important component, increasing from a low of 35 percent to 77 percent. The total levels of pay have risen six-fold.[3]

The shift of compensation practices toward options and restricted stocks has one immediate implication: the total compensation of the CEO is now

[3] The data is from *Forbes* magazine's annual survey. Prior to 1992, the data on options is a record of the gains from exercise. After 1992, *Forbes* records the value of the option grants at the date of granting, using the valuation method that the firm uses in its proxy statement. This method is mostly the Black and Scholes formula, for which the company reports the parameters as established in the SEC disclosure rules of 1993. See Section 4 of this article on regulation.

Also, note that good data on pensions is not available and most studies leave it out of their analysis.

Figure 2 Annual Compensation (*Forbes* Magazine) and Dow Jones Industrial Average During the Previous Year



more closely tied to the performance of the firm than it was in the 1980s. In Figure 2, I plot the *Forbes* measure of total compensation from Figure 1, together with the Dow Jones Industrial Average during the previous year, as an approximation for the average performance of the firms managed by the CEOs included in the compensation surveys. Apparently, on average, the compensation of top CEOs is closely correlated with the performance of the top firms that they manage. A more detailed analysis is needed to determine this relation at the individual firm level. In the next two sections, I review the empirical literature that has performed a detailed analysis, documenting the trends in the level and sensitivity of pay. In a widely cited paper, Jensen and Murphy (1990a) studied changes in compensation facts from 1974 to 1986. Following this paper, many studies on pay have been published. Rosen (1992) provided a first survey of the empirical findings on CEO compensation and Murphy (1999) is a very complete survey of the literature in the 1990s. The general picture, according to the most recent studies, is that the rates of growth in the level of CEO pay have increased dramatically since the 1980s, but pay has become, over the same period, more closely linked to firm performance. I present the evidence in two separate sections, one focusing on the level of pay and one on the sensitivity of pay—although, as will become clear

Table 1 Median CEO Pay

Period	Median CEO Pay (Million $)	Avg. Annual Growth Rate
1936–1939	1.11	
1940–1945	1.07	−0.01
1946–1949	0.90	−0.04
1950–1959	0.97	0.01
1960–1969	0.99	0.00
1970–1979	1.17	0.02
1980–1989	1.81	0.04
1990–1999	4.09	0.08
2000–2005	9.20	0.14

Source: Frydman and Saks (2007). CEO pay is in constant year 2000 dollars.

from the discussion, the two features are related through risk and incentive considerations.

2. THE LEVEL OF PAY

The bulk of the literature on CEO compensation (see Jensen, Murphy, and Wruck 2004; Frydman and Saks 2007; Gabaix and Landier 2008) has documented that, after moderate rates of growth in the 1970s, CEO pay rose much faster in the most recent decades. Total compensation reached a median value of $9.20 million (in 2000 constant dollars) in 2005. Frydman and Saks (2007) present previously unavailable historical evidence starting in the 1930s. They find that compensation decreased sharply after World War II and it had a modest rate of growth of about 0.8 percent for the following 30 years. Hence, the recent growth in pay (see Table 1 and Figure 1), with annual growth rates above 10 percent in the period 1998–2007, represents a remarkable change from previous trends.

In the first subsection, I review some estimations of the relevance and appropriateness of the levels of CEO pay that we observe today. I follow with a subsection on the proposed explanations for the time trends of the level of pay in the last century, which are divided into two groups: theories based on competitive markets and optimal contracts and theories based on inefficient practices.

Are Today's Levels of Pay Justified?

According to some recent studies, the cost of CEO compensation for firms may not be so economically significant. However, the economic consequences of not providing the right incentives appear to be potentially large. Gabaix and

Landier (2008), for example, report an average value of CEO pay over firm earnings of 0.5 percent during the period 1992–2003. Margiotta and Miller (2000), relying on a moral hazard model and historical data on compensation contracts of 34 firms from 1948 to 1977, find that companies benefit highly from providing the right level of incentives (ranging from $83 million to $263 million in 1977 dollars, depending on the industry). The observed contracts in their sample, however, imply a modest cost of providing incentives to induce high effort (about $0.5 million in 1967 dollars).[4] More recently, Bennedsen, Pérez-González, and Wolfenzon (2007) use data from Danish firms from 1992 to 2003 to estimate the effect of the CEO's death on firm performance. They find that the industry-adjusted operating returns on assets decrease by 1.7 percent following the absence of the CEO. They also report a decrease of 0.7 percent after a close family member of the CEO passes away. Both measures indicate the importance of CEO input into the performance of the firm, either through effort, talent, or both. Using a competitive model, Terviö (2008) estimates the talent distribution for CEOs and finds that if the managers of the 1,000 largest U.S. companies in 2004 had been substituted by a CEO of the same talent as that of the 1,000th company, the combined value of those companies would have been between $21.3 billion and $25 billion lower. His calculations imply that the contribution of the talented CEOs (net of pay) was between $16.9 billion and $20.6 billion (approximately 15 percent of the total market capitalization of the largest 1,000 firms), while the combined payments to the 1,000 CEOs that year were $7.1 billion.

Explanations for the Increase in the Level of Pay

The widely documented increase in the level and sensitivity of pay in the past two decades has motivated research efforts to construct robust explanations for these trends. Gabaix and Landier (2008) argue in a recent influential article that the six-fold increase in pay since the 1980s can be explained by the six-fold increase in the market value of U.S. firms in the same period. Building on work by Lucas (1978) and Rosen (1981, 1982), Gabaix and Landier's model presents a competitive economy in which firms compete for a manager's talent.[5] In their characterization of the equilibrium, they show that the wage of the CEO increases both with the size of the firm and the average size of the firm in the economy. Using extreme value theory, they calibrate the distribution for a manager's talent and combine this with a calibration of the size distribution of firms. Their results are consistent with the findings of Terviö (2008), who

[4] Margiotta and Miller (2000) use exponential utility specifications in their empirical estimations, for which incentive considerations are independent of the level of wealth.

[5] See Prescott (2003) for a related model.

independently used a competitive model to estimate the distribution of CEO talent that does not rely on extreme value theory.

Although the model in Gabaix and Landier (2008) can explain trends in CEO compensation since the 1980s, recent evidence in Frydman and Saks (2007) challenges their hypothesis for earlier periods. Frydman and Saks provide a comprehensive empirical account of the evolution of CEO pay in the last century. Their database includes previously unavailable data for the period 1936–1992, hand-collected from the 10-K filings of firms with the SEC. They study an unbalanced panel of 101 firms, selected for being the largest according to market value at three different dates. They find, consistent with previous work, that the increase in CEO pay was moderate in the 1970s and picked up in the last three decades. However, their new historical evidence starting in the 1930s shows that prior to the mid-1970s, while the level of pay was positively correlated with the relative size ranking of the firm in the economy, the correlation between changes in the market value of firms and the changes in the level of compensation was very weak. In particular, their data shows that CEO pay was stagnant from the 1940s until the 1970s, a period during which there was considerable growth of firms—particularly during the 1950s and 1960s.

Some authors in the literature have departed from the usual modeling approach by proposing that the observed contracts are in fact suboptimal. The following are three representative proposals of this sort.

Misperception of the Cost of Options

Murphy (1999) and Jensen, Murphy, and Wruck (2004) have argued that the significant increase in the use of option grants in recent years can be explained in part by the fact that compensation boards do not correctly valuate the cost of granting options. They argue that the tax and accounting advantages of options make it an attractive method of compensation.[6] They claim that, in fact, options are an expensive way of compensating the CEO: risk aversion, combined with the impossibility of hedging, as well as the high percentages of human and monetary wealth that CEOs invest in their own firm, and the positive probability of being unable to exercise the options if they are fired before the options become vested, all imply that managers ought to demand a high risk premium for the options. Thus, while firms are experiencing a cost that is well approximated by the Black-Scholes value of the grants disclosed in the proxy statements and listed in the executive compensation databases, CEOs highly discount their value. Hall and Murphy (2002) explicitly model these issues and provide calculations on how the CEO's risk aversion influences the

[6] Recent changes in regulation extend the expensing requirements to grants that have an exercise price equal to or above the market price at the time of granting.

valuation of the stock grants. They find that the increase of "risk-adjusted" pay from 1992 to 1999 has not been nearly as dramatic as that of total pay figures taken at face value.

Ratchet Effect

Murphy (1999) provides a description of how firms decide on the annual option grant to be awarded to executives. He points to the fact that firms typically use a measure of the average compensation given to executives in a peer group for reference. His data also indicates that 40 percent of firms have compensation plans that specify the number of options to be granted, fixed for several years, as opposed to the value of the options. In times of a growing stock market, such as the late 1990s, these two compensation practices together would result in a "rachet effect"—an escalation in total pay based on mutual benchmarking, as opposed to rewarding exceptional performance of the individual CEOs.

Entrenchment

Bebchuck and several coauthors (see, for example, Bebchuck and Fried [2003, 2004] and references therein) have argued that there is ample evidence suggesting that executives control the boards of directors and effectively determine their own pay. Kuhnen and Zwiebel (2008) provide a formal model of this hypothesis, which presents the problem of the manager maximizing his own pay subject to an "acceptable" outrage level of the shareholders. The mainstream literature accepts that some incentive problems remain unsolved given the current compensation and corporate practices in effect (see Holmström and Kaplan 2003 or Jensen, Murphy, and Wruck 2004). Yet, there is no consensus on whether the entrenchment model is a better description of real life than the classic moral hazard model, which has shareholders maximizing their value subject to the incentive constraints of CEOs (see Grossman and Hart 1983). It is also worth noticing that the entrenchment explanation per se does not help in understanding the upward trend in the level of pay observed in the last 15 years, since corporate governance practices have been improving during this period (see Holmström and Kaplan 2003).

3. THE SENSITIVITY OF PAY TO PERFORMANCE

The literature on CEO compensation uses the term "sensitivity of pay to performance" to refer to the changes in the total pay of a CEO that are implied by a given measure of the performance of the firm. The most common measure of firm performance used in the empirical estimations of the sensitivity of pay is the change in the value to shareholders, i.e., stock price change. Theoretical studies based on moral hazard models have shown that the design of the optimal compensation scheme is a complicated task. The way in which the level

of risk aversion of undiversified CEOs decreases with their wealth (which may originate both from their firm pay and from outside sources), for example, has not been well established empirically (Haubrich 1994). Also, finding empirical evidence on the parameters of the model is difficult: How does the hidden action of the CEO affect the results of the firm and how costly is this action to the CEO? These considerations make it difficult to assess quantitatively the optimal sensitivity of direct pay. Moreover, the sensitivity itself is difficult to estimate empirically, since the pay of the CEO changes with firm performance both through direct and indirect channels: The competition for talented CEOs in the market, for example, implies that career concerns and the risk of being fired impose some incentives on the executives that are not captured by their compensation contracts.

In the first subsection, I review the academic work that attempts to quantify the incentives provided through direct pay. In the second subsection, I review the work that quantifies incentives through indirect pay, consisting mainly of studies that document historical trends of the probability of employment termination for CEOs. Although there is no agreement in the literature about the right measure of sensitivity, two stylized facts are widely accepted: total (direct plus indirect) sensitivity of pay has been increasing in the last two decades and sensitivity is negatively correlated with firm size. In the third subsection, I review some of the proposed explanations for these stylized facts, as well as several models that justify the diversity of instruments in real life compensation packages.

Sensitivity of Direct Pay

In an influential paper, Jensen and Murphy (1990a) point out the seemingly low sensitivity of executive pay to performance: an increase in total wealth of \$3.25 for each \$1,000 of increase in the firm value. This measure of sensitivity is, in the data, highly negatively correlated with the size of the firm. As an alternative that allows for variation of the Jensen and Murphy (1990a) measure across firm size, other studies have estimated the elasticity of pay to firm performance. For both measures, a sharp increase in the sensitivity of pay over the last two decades has been widely documented: estimations suggest an increase in sensitivity of more than five times over that period.

I now describe in detail each of the two measures and the main findings, including a discussion of the regularities with respect to firm size.

The Jensen-Murphy Statistic

Jensen and Murphy (1990a) use data on compensation of approximately 2,000 CEOs of large corporations collected by *Forbes* magazine from 1974 to 1986. In their paper, they focus on a particular measure of pay sensitivity to

performance: They estimate how a $1,000 change in the value of a firm translates into dollar changes in the wealth of its CEO. This measure of sensitivity is often reported in the literature as the Jensen-Murphy statistic.

In their paper, they run regressions of the form

$$\Delta w = \alpha + \beta\left(\Delta V\right) + \theta\left(CONTROLS\right), \tag{1}$$

where ΔV includes lagged as well as contemporaneous measures of changes in the dollar value of the firm. They report that the median CEO in their sample experiences an increase in his total wealth of $3.25 for each $1,000 of increase in the value of the firm he manages. Of this change in wealth, 2 cents correspond to year-to-year changes in salary and bonus, while the rest is because of changes in the value of the stock and option holdings of the CEO, the effect of performance on future wages and bonus payments, and the wealth losses associated with variations in the probability of dismissal. The authors qualify this sensitivity of pay as "low" and inconsistent with the predictions of any agency model. Murphy (1999) updates these estimates using Execucomp data up to 1996 and reports an increase in the sensitivity of pay to $6 per $1,000 by that year.

Garen (1994) and Haubrich (1994) point out that the seemingly low estimates of Jensen and Murphy are consistent with optimal incentives in the presence of moral hazard if the right parameters of risk aversion are chosen for the specific functional forms. Garen (1994) presents a static moral hazard model with closed forms solutions and, as an alternative test for the theory, derives comparative static predictions, which do not rely on the particular values of parameters and are more robust to functional form specifications. He concludes that most implications of moral hazard are consistent with the data. In the same spirit, Wang (1997) provides a computed solution to a dynamic agency (repeated moral hazard) model that, under a reasonable range of parameters, is able to replicate the low sensitivity of pay results in Jensen and Murphy (1990a). The model shows that deferred compensation is sometimes a more efficient way of providing incentives than current pay because of incentive-smoothing over time.

Changes in Sensitivities with Firm Size

The Jensen-Murphy statistic has been documented to be very different across different firm size subsamples: Jensen and Murphy (1990a) divide their sample according to market value and find a total pay sensitivity value of $1.85 for the subsample of larger firms versus $8.05 for that of smaller firms.[7] Updates of this measure provided in Murphy (1999) suggest that the increase in median pay sensitivity from 1992 to 1996 was a lot higher for large firms than

[7] See, also, Jensen and Murphy (1990b).

for smaller ones (64 percent [from \$2.65 to \$4.36] for the largest half of the S&P 500, 5 percent [from \$7.33 to \$7.69] for the smallest half of the S&P 500, 28 percent [from \$12.04 to \$15.38] for the S&P Mid-Cap corporations, and 24 percent [from \$22.84 to \$28.23] for the S&P Small-Cap corporations). Garen (1994) also finds a negative relationship between sensitivity of pay and firm size. Schaefer (1998) finds that the Jensen and Murphy (1990a) measure is approximately inversely proportional to the square root of firm size. Recent estimations of this measure use quantile regressions to prevent estimations from being driven by big firms in the samples.[8]

The literature agrees that the Jensen-Murphy measure is adequate for firms in which the marginal effect of effort is independent of firm size. However, it has been repeatedly pointed out that it does not correctly capture incentives when the marginal product increases with firm size. Baker and Hall (2004) propose a model that relates compensation to the marginal product of the CEO. They confirm the previous estimates of decreasing sensitivity of pay to firm size and they identify higher marginal products of CEOs working in bigger firms as the main cause.[9] Their model implies that, although the Jensen and Murphy (1990a) measure of sensitivity falls sharply with firm size, the total incentives of CEOs remain constant or fall only slightly since the effect of CEOs' actions increases with firm size. A similar assumption is used in the competitive market model of Edmans, Gabaix, and Landier (forthcoming), in which the effort cost for the CEO is bounded above, but his marginal impact depends on the size of the firm he manages.

Elasticity of Pay to Firm Value

An alternative measure of sensitivity of pay is the elasticity of compensation to performance: the percent increase in executive wealth for a 1 percent improvement in the performance of the firm. Note that assuming a constant elasticity across firms implies variation in the Jensen-Murphy statistic across firm sizes.

A convenient way of estimating the elasticity of pay to firm performance is to regress the logarithm of earnings on the return of the firm's stock:

$$\ln w_t = a + \widetilde{b} \times r_t. \tag{2}$$

Then $\widetilde{b} = d(\ln w)/dr$ measures the semi-elasticity of earnings to firm value and the elasticity is recovered for each particular return value as approximately $b = r\widetilde{b}$. Initial estimates of the semi-elasticity of cash compensation (salary and bonus) to stock returns, surveyed in Rosen (1992), had reported

[8] See, for example, Murphy (1999) and Frydman and Saks (2007).

[9] This is consistent with the hypothesis that more talented CEOs work for bigger firms, as mentioned in Rosen (1992). See, also, the competitive market for talent models of Gabaix and Landier (2008) and Terviö (2008) reviewed later in this article.

a median value for b of about 0.10—an increase in a firm's return of 10 percent ($\Delta r = 0.1$) implies a 1 percent increase in pay.[10] Using average values for w and r, Rosen (1992) reports that these semi-elasticity values imply a Jensen-Murphy statistic for cash compensation (salary plus bonus) of 10 cents, compared with their finding of 2 cents.[11] He concludes that lower sensitivities of pay at bigger firms probably influence the Jensen-Murphy estimation, especially since they do not log the compensation figures.

Hall and Liebman (1998) provide estimates of the elasticity of compensation using detailed data on CEO pay from 1980 to 1994. To correctly estimate the elasticity and several other measures of sensitivity of pay (including the Jensen-Murphy statistic), they collect data on stock and options grants from the proxy statements of firms. They construct the portfolio of stock and options for each CEO in their databases in 1994: This way they can include in their measure of total pay the variation in the wealth of the CEO because of the changes in the price of the firm's stock.[12] Also, they can evaluate the potential changes in wealth for different realizations of the stock price (the ex ante sensitivity of pay). Since changes in the wealth of the CEO are sometimes negative because of the decrease in the value of their stock holdings, Hall and Liebman cannot directly run the regression in (2) to calculate the semi-elasticity. Instead, with information both on current total compensation as a function of the firm's return in 1994, $w(r_t)$, and on the distribution of annual stock returns, they predict changes in compensation for a given change in firm return, a hypothetical $\widetilde{w}(r_{t+1})$. With this predicted total compensation measure, they calculate the semi-elasticity directly from the formula

$$\frac{\widetilde{w}(r_{t+1}) - w(r_t)}{w(r_t)} = \widetilde{b} \times (r_{t+1} - r_t).$$

In their data for 1994, Hall and Liebman evaluate the effects of a typical change in firm return: from a median performance ($r = 5.9$ percent) to a 70th percentile performance ($r = 20.5$ percent), which implies a 14.6 increase in r.[13] For this typical change, they calculate the semi-elasticity for each CEO

[10] Rosen (1992) points out the discrepancy of the findings in the literature with those of Jensen and Murphy. The sensitivity of salary and bonus in Jensen and Murphy (1990a) is 1.35 cents for $1,000. This number implies an elasticity at the mean firm size of 0.0325; equivalently, an elasticity of 0.1 translates into a change of 10 cents per $1,000 increase in firm value. Rosen attributes these differences to functional forms and to the sensitivity measure, which is dominated by the large firms' observations, with significantly lower sensitivities.

[11] Rosen (1992) uses an average value for w/V of 10^{-3}. Following his calculations, the Jensen-Murphy statistic from equation (1) is $\beta \approx b\frac{w}{V}$.

[12] Jensen and Murphy (1990a) also include in their total compensation measures the changes in the wealth of the CEO attributable to his holdings of stocks and options. Their sample, however, ends in 1983, before the significant increase in option grants that is captured in the Hall and Liebman (1998) sample up to 1994. Moreover, the sample in Jensen and Murphy (1990a) is from *Forbes* 800 and includes the value of exercised options as opposed to the value of options at the time of granting. After 1992, Execucomp started collecting this information.

[13] The median standard deviation of r is about 32 percent in their data.

in their data set:

$$\widetilde{b} = \frac{\frac{\widetilde{w}(0.205) - w(0.059)}{w(0.059)}}{0.146}.$$

The implied mean elasticity is 4.9 and the median is 3.9. Hall and Liebman also compute the Jensen-Murphy median sensitivity based on variation in stock and options only: they find a value of \$6, compared to \$3.25 reported in Jensen and Murphy (1990a) for the period 1974–1986.[14] They confirm the finding in Jensen and Murphy (1990a) that incentives are provided mainly through the granting of stocks and stock options. The sensitivity of salary and bonuses to firm performance is very low, while the changes in the wealth of the CEO that originate from his stock and option holdings are very big. They find a large increase in the use of option grants in the period covered by their sample, which translates into a significant increase in the sensitivity of various pay to performance measures: Between 1980 and 1994, the median elasticity went from 1.2 to 3.9, while the median wealth change per \$1,000 (the Jensen-Murphy statistic) went from \$2.5 to \$5.3.

Indirect Sensitivity: Provision of Incentives through Career Concerns or Threat of Dismissal

Even if the total pay of a CEO was independent from the performance of his firm, the manager still would have some incentives to exert effort if the threat of dismissal was high enough or career concerns were present. Career concerns is the term used to summarize the fact that workers expect to receive offers for better paying jobs after an above-average performance. Several studies have tried to quantify the importance of these two implicit incentive channels.

Jensen and Murphy (1990a) recognize the existence of indirect provision of incentives that is implicit in the threat of dismissal of the CEO following poor performance. To account for those indirect incentives in their sensitivity measure, Jensen and Murphy provide an approximation of the wealth variations for the CEO that would follow if he were fired from his job. To estimate this wealth loss, they first estimate the probability of CEO turnover as a function of firm performance (net-of-market return in the current and previous year). They find that a CEO in a firm with average market returns in the past two years has a .111 dismissal probability, while a performance 50 percent below market returns for two years increases the dismissal probability to .175. Since they do not have information on whether the separation was voluntary

[14] Hall and Liebman also report the mean sensitivity of pay in their sample, equal to \$25 without evaluating potential changes in wealth due to the threat of dismissal. They claim that the large differences between the mean and the median values are due to the high skewness of stock and option holdings in the sample of CEOs. Jensen and Murphy (1990a) do not report mean values.

or because of retirement, these estimated values represent a rough measure of the real probability of dismissal. They also find evidence of greater effects of performance on turnover for younger CEOs and for smaller firms. They use these estimated probabilities and a simplifying assumption that a fired executive will earn a salary of $1 million elsewhere until retirement to calculate the dismissal-related wealth loss of CEOs of various ages and as a function of their net-of-market return. For example, a 62-year-old CEO would suffer a loss of 26.4 cents for every $1,000 lost by shareholders if his firm performed 75 percent below the market, as opposed to performing the same as the market.[15] The highest sensitivity they find is for younger CEOs, who would experience a loss of 89 cents per $1,000.

Subsequent studies have found an increase in job turnover probabilities in recent years, as well as evidence of the importance of relative performance as a determinant of firing decisions. Kaplan and Minton (2006) extend the analysis for the period 1992–2005, and they find an increase in the probability of turnover with respect to previous periods. They report a decrease of the average tenure of a CEO from over ten years, as reported in Jensen and Murphy (1990a), to less than seven years for the period 1992–2003 (which corresponds to a probability of turnover of 0.149). The average for 1998–2003 is significantly lower, at just over six years (a probability of 0.165), and this subperiod shows higher sensitivity to current measures of performance than in previous periods. Kaplan and Minton include in their analysis three measures of performance: firm performance relative to industry, industry performance relative to the overall market, and the performance of the overall stock market. They find that turnover initiated by boards is sensitive not only to firm performance measures but also to bad industry or economy-wide performance. They interpret this fact as indicative of a change in corporate governance since the 1970s and 1980s, when bad industry or economy-wide performance influenced CEO turnover in the form of hostile takeovers.

In a related study, Jenter and Kanaan (2006), using a new sample including both voluntary and involuntary turnovers during 1993 to 2001, confirm the results in Kaplan and Minton (2006). They find that both poor industry-wide performance and poor market-wide performance significantly increase the probability of a CEO being fired. A decline in the industry performance from the 75th to the 25th percentile, for example, increases the probability of a forced CEO turnover by approximately 50 percent, controlling for firm-specific performance. The firm-specific performance measures are also weighted by boards more heavily when the overall market and industry performance is worse.

[15] 1986 constant dollars.

Gibbons and Murphy (1992) derive a model with explicit contracts and career concerns that builds on Holmström (1999). Career concerns represent the expectation over future wages based on the market beliefs about the quality of the CEO. They present empirical evidence that explicit incentives increase as the CEO gets closer to retirement age: for one, two, or three years left in office, respectively, they find elasticities of pay to performance of .178, .203, and .183, compared with elasticities of .119 and .116 when they have five or six years left. They interpret this evidence as support for their theoretical findings that career concerns complement the explicit incentives provided by the ties of current wealth to performance in their current employment.

Overall, the available measures of indirect sensitivity of pay to performance appear to have been increasing in the last three decades, in accord with the increase in the direct sensitivity of pay.

Explanations of Sensitivity Facts

Two main empirical regularities emerge from the studies reviewed previously. First, the sensitivity of pay is smaller for CEOs of larger firms. Second, the sensitivity of pay across all firm sizes has increased in the last two decades. In the following subsections, I review the main theoretical explanations proposed for these two facts. Finally, I include a subsection that presents justifications for a set of characteristics of real life compensation contracts. These characteristics may, to the uninformed eye, seem unappealing for the provision of incentives. The theoretical models show, however, that they may just be the optimal instruments to implement sensitivity of pay to performance.

Fact 1: The Sensitivity of Pay Decreases with Firm Size

Baker and Hall (2004) show that, in the presence of moral hazard, the sensitivity of pay optimally decreases with the size of the firm. They rely on a partial equilibrium model with multitasking and heterogeneity in the marginal productivity of the CEO as a function of the firm size. In recent work, Edmans, Gabaix, and Landier (forthcoming) extend Gabaix and Landier (2008) and present an agency model embedded in a competitive market model for talent. They propose a measure of sensitivity of pay that their theory predicts as independent of firm size: the dollar change in wealth for a percentage change in firm value, scaled by annual pay.

Fact 2: There has been an Increase in Sensitivity of Pay in the Last Two Decades

The explanations for the increase in sensitivity of pay to performance fall mostly into two categories: those that maintain that the increase has been driven by changes in firm characteristics and increased market competition

and those that maintain that the increase has been driven by an improvement in corporate governance practices.

A good example of the first category of explanations is the work of Cuñat and Guadalupe who, in a series of three papers (2004, 2005, 2006), evaluate the changes in CEO compensation following several changes in market conditions:[16] Cuñat and Guadalupe (2004) study two episodes of deregulation of the U.S. banking system and find that the increase in competition is correlated with an increase in sensitivity of pay; Cuñat and Guadalupe (2005) find that in the period following the sudden appreciation of the pound in 1996, which implied different levels of competition for tradables and nontradables, we observe an increase in sensitivity of pay for executives in the tradables sector; Cuñat and Guadalupe (2006) find, for a sample of U.S. executives, that industries more exposed to globalization (higher foreign competition as instrumented by tariffs and exchange rates) exhibit higher sensitivity of pay, higher dispersion of wages across different levels of executives, and higher demand for talent at the top, which drives up the salary of the CEO. This evidence supports the conclusions of some theoretical papers that analyze the effects of increased competition for talented CEOs. Frydman (2005), for example, presents a dynamic model of executive compensation and promotions in which the superior information of firms about their incumbent workers implies that an increase in the relative importance of general skills versus firm-specific skills triggers an increase in manager turnover, increased differences in pay among executives of different ranks of the same firm, and higher levels of pay. Frydman contrasts these implications with historical data on compensation, careers, and education of top U.S. executives from 1936 to 2003. She concludes that facts are consistent with the predictions of her model under the hypothesis of an increase in the relative demand of general skills after the 1970s.

The work of Holmström and Kaplan (2003) falls under the second (possibly complementary) category of explanations. They argue that the changes in corporate governance of U.S. firms in the last two decades have had a net positive impact, translating into higher sensitivity of pay to performance. Although they recognize that the increase in options compensation has probably led to new incentive problems that may be behind the corporate scandals of recent years, they argue that the evidence is still in favor of an improvement in the management of firms. They identify three main positive signs: the good comparative performance of the U.S. stock market with respect to the rest of the world, the undisputable increase in the link between CEO compensation and firm performance, and the use of compensation schemes by buyout investors and venture capital investors that resemble those of the average public firm. They review the trends in corporate governance since the

[16] See, also, references in Gabaix and Landier (2008).

1980s and they identify, as a sign of success, the increasing convergence of international governance practices to those of the United States, which were drastically different from those of, for example, Germany and Japan during the 1980s. They also point to an increase in shareholder activism, which may be due to an increased institutional shareholders' stake in public companies: their overall share in the total stock market went from 30 percent in 1980 to more than 50 percent in 1994. Gompers and Metrick (2001) report evidence supporting the hypothesis in Holmström and Kaplan (2003) that institutional investors improve corporate governance. They find that over the period 1980–1996, firms with higher shares of institutional shareholders had significantly higher stock returns. Gompers, Ishii, and Metrick (2003) construct an index to proxy for the level of good governance of firms based on data on completion of a number of governance provisions aimed at protecting shareholder's rights. They collect data on four dates between 1990 and 1998 and include a very high proportion of the largest U.S. firms according to market capitalization. They find strong correlation between good governance and good firm performance in terms of Tobin's Q (market value over book value), profits, and sales growth. Holmström and Kaplan (2003) also identify signs of improved board of directors' independence: an increase in CEO turnover, a decrease in the size of boards, and an increase in the proportion of the compensation of board members that is in the form of stock and options of the firm.[17] An earlier theoretical paper, Hermalin and Weisback (1998), provides support for the importance of board independence on sensitivity of pay and, in general, CEO performance. This paper models explicitly the determination of independence levels of the board of directors. The CEO uses his perceived higher ability with respect to any alternative hire to bargain with the board and influence its composition. Their model has implications consistent with empirical facts for the 1980s and 1990s, such as CEO turnover being negatively related to prior performance, with greater effect for firms with more independent boards and with accounting measures being better predictors of turnover than stock price performance. Also, the independence of the board is likely to increase after poor firm performance and decrease over a CEO's career.

Diversity of Compensation Practices

In spite of the existence of some stylized facts on sensitivity of pay, real life compensation packages are highly complex and diverse. It is easy to find examples of features of compensation instruments and practices that may seem counterintuitive to the uneducated eye, such as the lack of relative

[17] Compensation boards are independent committees that design the compensation of the top managers of the firm. See Section 4 on regulation for details on the legal requirements for defining this board and the board of directors as "independent."

performance evaluation, or the repricing of options gone out-of-the-money. However, changing environments and dynamic considerations sometimes imply that such features are part of an efficient provision of incentives. What follows is a brief (nonexhaustive) list of recent theoretical articles that provide justification for some controversial compensation practices.

Commitment. Clementi, Cooley, and Wang (2006) show in a dynamic model of CEO compensation that implementing optimal incentives for executives in the presence of moral hazard and limited commitment through the issue of securities dominates deferred cash compensation arrangements; granting securities improves commitment and helps retain CEOs.

Dynamic Incentives. Wang (1997), as discussed previously, provides a model of repeated moral hazard that can explain very small (or even negative) sensitivities of CEO pay based on the optimal smoothing of incentives over time.

Learning. Celentani and Loveira (2006) show how learning about common shocks in the economy can explain the apparent absence of relative performance evaluation documented in the literature.[18] In the presence of moral hazard, if the productivity of the CEO's effort depends on aggregate conditions, the optimal compensation contract is not necessarily decreasing in poor relative performance.

Short-term Stock Price Performance. Current compensation plans are often criticized for providing incentives for CEOs to engage in actions that increase stock prices in the short term as opposed to creating long-term value for the shareholders.[19] Bolton, Scheinkman, and Xiong (2006) present a model that explicitly accounts for liquidity needs as a reason to participate in the stock market. They show that shareholders may in some circumstances benefit from artificial increases in today's stock prices, which influence the belief of speculators and allow the firm to sell its stock at higher prices in the future. This implies that compensation packages are sometimes optimally designed to make CEOs take actions that make short-term profits higher.

Targeting Efforts of the CEO. Kadan and Swinkels (2008) present a model in which the choice between restricted stock grants or option grants is linked to the potential effect of a manager's actions on firm survival. In financially distressed firms or startups where this effect is higher, stock is more efficient than options. The choice of compensation instruments helps the firm tailor the sensitivity of pay to the firm's idiosyncratic position. They find empirical evidence of stock being more widely used than options in firms with a higher probability of bankruptcy.

Repricing. The public perception of the practice of decreasing the exercise price of options that are out-of-the-money is that it "undoes" incentives

[18] See, for example, Gibbons and Murphy (1990).

[19] See, for example, Jensen, Murphy, and Wruck (2004).

and is thus interpreted as a sign of management entrenchment. This practice of repricing is fairly uncommon: Brenner, Sundaram, and Yermack (2000) report that 1.3 percent of the executives in their sample (top five officers in a sample of 1,500 firms between 1992 and 1995) had options repriced in a given year. However, it has received academic attention, perhaps motivated by its reputation as a bad compensation practice. Chance, Kumar, and Todd (2000) identify size as the main predictor for firm reprices, with smaller firms repricing more often. Brenner, Sundaram, and Yermack (2000) find that higher volatility also significantly raises the probability of repricing. Carter and Lynch (2001) find that young high technology firms and those whose outstanding options are more out-of-the-money are more likely to reprice. Chen (2004) finds that firms that restrict repricing have a higher probability of losing their CEO after a decline in their stock price and that they typically grant new options in those circumstances, possibly in an effort to retain the CEO. Acharya, John, and Sundaram (2000) present a theoretical model that implies that the practice of repricing can be optimal in a wide range of circumstances. The intuition is that there are benefits to adjusting incentives to information that becomes available after the initial granting and before the expiration of the options; in many cases, these benefits offset any loss in ex ante incentive provisions because of the lack of credibility from possible repricings.

4. REGULATION

The current levels of CEO compensation are viewed by many as excessive and unjustified. It is not uncommon to see demands in the popular press for government regulation of CEO pay.[20] However, the effects of regulation in a complicated matter such as the design of compensation packages for top executives are not always clear. An example of this is the change in regulation that was introduced in 1993 that limits tax deductions for any compensation above $1 million. This limit was introduced partly as a response to the popular rejection of the big increases in CEO pay that took place in the early 1990s. The law, however, established an exception to the limit: "performance-based" compensation (such as, for example, bonuses and options granted at the money).[21] Empirical studies cited below have found that firms have shifted compensation away from salary and toward stocks and options in response to the limit. A popular view today is that the inclusion

[20] As a recent example, the Emergency Economic Stabilization Act of 2008 passed by Congress on October 3, 2008, explicitly includes some limits on CEO compensation for firms in which the Treasury acquires assets or is given a meaningful equity or debt position. The compensation practices of such firms "must observe standards limiting incentives, allowing clawback and prohibiting golden parachutes." Restrictions also apply to firms that sell more than $300 million in assets to the Treasury through auction.

[21] For references, see United States Senate (2006).

of stocks and options in the compensation packages, together with the stock market boom, is partly responsible for the recent increase in CEO pay. Did the $1 million limit contribute sizably to the increase in pay by encouraging the use of performance-based compensation? In this section, I review the academic studies that have studied this question, as well as the effect of other changes in the tax treatment of CEO compensation.

Regulation has also been introduced in recent years to improve corporate governance in the United States.[22] Mainly, the measures have targeted improved coordination and power of shareholders, as well as the transparency of compensation practices. As a recent example, in the aftermath of corporate accounting scandals at Enron, WorldCom, and other important U.S. firms, the Sarbanes-Oxley Act of 2002 (SOX) increased the legal responsibilities of CEOs and of compensation committees. There are academic studies on the effect of regulation on corporate governance and I summarize their main conclusions in the second part of this section.

Tax Advantages

Tax exemptions for deferred compensation have been in place since the early 1980s. These tax exemptions apply to the capital gains tax owed by a CEO, which is levied for capital gains that are part of his compensation. We may wonder whether the savings on capital gains tax for stock option grants could have played a role in explaining the spectacular increase in the use of options in compensation packages. Also, some of the regulatory efforts that have stated specific limits for tax deductions have triggered criticisms since firms that were paying less to their executives may now raise their salaries to the limit, which is rendered by the regulation as acceptable. For example, with the passing of the Deficit Reduction Act of 1984, the U.S. government imposed a tax on excessive (three times the executive's recent average remuneration) "golden parachutes," i.e., payments to incumbent managers who were being fired in relation to a change in control. Jensen, Murphy, and Wruck (2004) argue that these types of agreements were fairly uncommon before the law was passed and that the regulation had the effect of endorsing the practice, which became a standard in the following years. A similar argument has been made for the Omnibus Budget Reconciliation Act resolution 162(m) of 1992, which imposed a $1 million cap on the amount of the CEO's nonperformance-based compensation that qualifies for a tax deduction.

In this section, I review a few of the academic studies that have tried to quantify the effect of tax advantages on CEO compensation. The main conclusion is that there exists a tax advantage of deferred compensation, although

[22] See Weinberg (2003) for a discussion of the financial reporting issues in corporate governance and of the Sarbanes-Oxley Act of 2002.

a modest one. Also, studies of the effect of resolution 162(m) have found evidence of a certain increase in salaries of companies that were paying less than $1 million before the new law was enacted, as well as some shifting of compensation toward performance-based instruments and away from salaries. There is evidence that this shift translated into a slight increase in total compensation. The studies, however, do not find evidence that the majority of the spectacular rises in the level of compensation and the increases in the use of options can be attributed to tax reasons.

Quantifying the Effect of the Special Tax Treatment for Compensation in the Form of Options

Currently, pay in the form of options has a tax advantage for CEOs when compared to payments in salary: There are no capital gains tax charges on the increase in value of a stock from the grant date to the exercise date. If, after exercise, the executive holds the stock and there is further increase in its value, capital gains tax is owed only on the appreciation from the date of exercise to the date of sale.

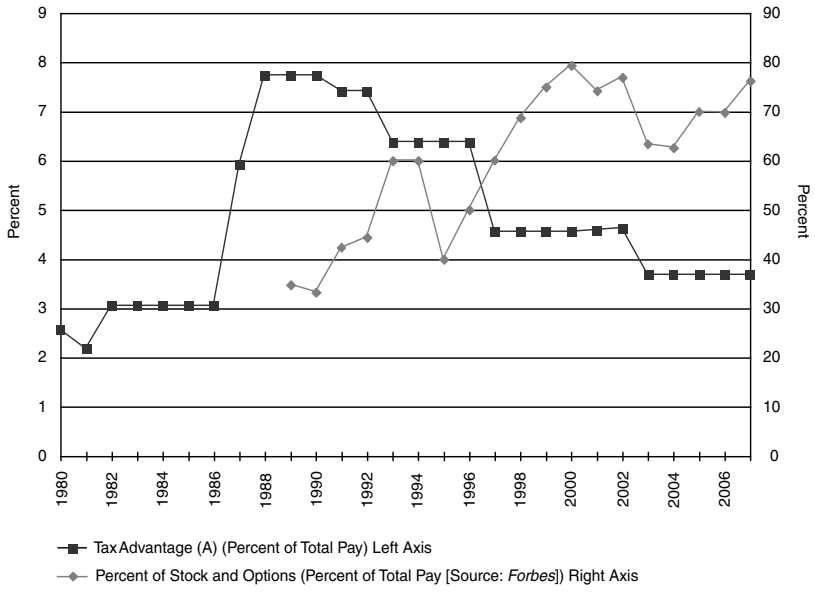
Motivated by different tax treatments of various instruments of compensation, Miller and Scholes (2002) explicitly compare the after-tax value (together for the firm and the CEO) of compensation in wages and several deferred compensation instruments (deferred salary schemes, stock purchase plans, deferred stock plans, nonqualified option plans, and insurance plans). Miller and Scholes document that by 1980 there had been already a rapid increase in the usage of deferred compensation versus bonus and salary. The use of these instruments is usually interpreted as a sign of incentive provision: With the granting of stock options, the compensation board attempts to align the incentives of the CEO with those of the shareholders, which helps increase the long-term value of the company. However, they find that none of the instruments they analyze have clear tax disadvantages with respect to wage payments: they are all either beneficial or tax-neutral. They argue that the tax savings makes it difficult to identify incentive provision as the only reason for deferred compensation.

The advantage of options versus pay is also analyzed in Hall and Liebman (2000) with data for 1980–1994.[23] A simplified version of their analysis is presented here. We must compare two alternative ways of compensating the CEO, both with net present value of P for the firm. On one hand, the firm could grant options to the CEO for a present value of P, which would vest in N years. The return of this compensation to the CEO, assuming an annual return of r for his firm, is

$$P\left[1 + r\left(1 - T_c\right)\right]^N \left(1 - T_p\right),$$

[23] For details on current tax treatment of CEO compensation, see United States Senate (2006).

Figure 3 Evolution of Tax Advantage Compared with the Proportion of Pay Given in the Form of Stock and Options



Source: IRS and *Forbes* magazine.

where T_p, T_c, and T_{cg} are, respectively, the personal, corporate, and capital gains tax rates. On the other hand, the firm could schedule a cash payment of P today. Assuming the CEO invests the whole after-income tax amount in a nondividend-paying security with the same return, r, for the same number of years, N, the return of the cash payment to the CEO is

$$P\left(1-T_p\right)\left[1+r\left(1-T_c\right)\right]^N - T_{cg}P\left(1-T_p\right)\left\{\left[\left(1+r\left(1-T_c\right)\right)^N - 1\right]\right\}.$$

Therefore, the options advantage, A, depends on the three tax rates:

$$A\left(T_p, T_c, T_{cg}\right) = T_{cg}P\left(1-T_p\right)\left\{\left[\left(1+r\left(1-T_c\right)\right)^N - 1\right]\right\}.$$

Hall and Liebman (2000) report the value of A for the period 1980–1998 by substituting the corresponding rates, keeping everything else equal. They find that the advantage of options versus cash compensation reached a maximum of \$7 per \$100 of compensation P in 1988–1990 and was down to \$4 per \$100 in 1998. Figure 3 plots the value of the advantage reported in the paper, plus the value for the period 1980–2007 using recent tax rates. The Bush tax breaks have decreased the advantage of options in recent years;

after they expire, and with no further changes, the advantage should go back to the \$4 calculated for 1998. In the same figure, on the right axis, I have plotted the ratio of option grants over total compensation from 1992 to 2007 from the *Forbes* data previously shown. Comparing the two series, we can see that, although tax advantages were decreasing in the later period, options were being used relatively more intensively than salaries. Hall and Liebman (2000) complement their analysis with regressions of the use of options on the several tax rates and their calculation of the tax advantage and find no evidence of a sizable effect.

Quantifying the Effect of Caps on Deductibility of CEO Salaries

When firms calculate their corporate tax liability, under the Internal Revenue Code, they are allowed to subtract from their revenues the compensation of their workers. Prior to 1993, this deduction included all compensation given to the top executives of the firm.[24] In 1993, new regulation took effect that limited the deductions of the compensation of each of the five highest ranked executives of a firm to \$1 million (section 162(m) of the Internal Revenue Code). The regulation specifies an exception to the \$1 million limit: Compensation that is performance-based can be deducted regardless of being worth more than \$1 million.

The requirements to qualify as performance-based for each form of compensation can be summarized as follows:

- Salaries are, by nature, nonperformance-based. Any company paying its CEO a salary higher than \$1 million would face higher tax costs after the change in regulation.

- A bonus program qualifies if it specifies payments contingent on objective performance measures and is approved in advance by the vote of the shareholders.[25]

- Stock options are considered entirely performance-based compensation. In order to qualify for tax deduction, firms have to detail in an option grant plan the maximum number of options that can be awarded

[24] The accounting of compensation expenses is briefly discussed later in relation to regulation of disclosure of information to shareholders and the valuation of options.

[25] Murphy (1999) documents that companies do include subjective measures of the performance of the CEO in their bonus plans, usually labeled "individual performance." In the typical bonus plan in his data source, this subjective component rarely exceeds 25 percent of the executive's bonus.

to a given executive in the specified time frame. The plan needs to be approved by the shareholders.[26]

- Stock grants and other types of deferred compensation (pensions, severance payments, etc.) are only considered performance-based if they are granted or their selling restrictions vest based on performance targets.

I now review the three main academic studies of the effect of this change in regulation.

A simple strategy of testing for the effect of section 162(m) on pay practices is to regress changes in compensation on changes in firm value:

$$\Delta \ln (w_{it}) = \beta_0 \Delta r_{it} + \delta_t + \varepsilon_{it},$$

where δ_t captures the annual growth rates.[27] Lower values of δ after the reform would imply the regulation succeeded in slowing the growth of executive compensation. However, the trends in CEO compensation may be responding to changes in the environment other than regulation. To identify the effect of the reform, the empirical studies have exploited the cross-sectional variation in the population of firms. By separating the firms in which compensation practices were not affected by the cap on deductions, they can use those as a control group and isolate the effect of the regulation:

$$\Delta \ln (w_{it}) = \beta_0 \Delta r_{it} + \alpha_t AFFECTED + \delta_t + \varepsilon_{it},$$

where the difference in α_t before and after the change in regulation captures the effect of deduction limits on the growth of pay in the affected firms.[28] Deciding which firms should be classified as affected is not such a simple task, however. Hall and Liebman (2000), for this purpose, use data on the compensation in 1992, the only observation available in the Execucomp data set before the change in the law. They construct a variable equal to 1 if the salary observation in 1992 is above $1 million, and equal to $1 million divided by the salary for firms below that level. They find that a CEO with a $1 million salary in 1992 saw, after the new regulation, his salary grow at an annual rate of approximately 0.6 percent less than an executive earning $500,000.

Perry and Zenner (2001) propose three alternatives for identifying the set of firms affected by the regulation. One is the same indicator constructed in

[26] These plans are usually specified for several years. If, after approval, the compensation board finds it necessary to exceed the total number of options specified in the plan for a given CEO, the excess amount of options does not qualify for a deduction.

[27] In the empirical studies, several measures of performance (sometimes logged, sometimes in differences) and lagged performance are included as explanatory variables, as well as other controls like CEO tenure and other relevant available information. In this section, I use a simplified statement of the regression equations meant only for illustration.

[28] It has also been argued that setting the deduction cap at $1 million encouraged firms that were far below this level of compensation to increase their manager's pay, since $1 million became "acceptable" (see *Wall Street Journal* 2008). If this effect really existed, it would bias the estimates in the above regression.

Hall and Liebman (2000), a second one is an indicator that takes a value of 1 if the salary in 1992 is above $900,000, and a third is an indicator that takes a value of 1 if the executive had a combined salary and bonus payment above $1 million in any year before the change in regulation. They find similar results in the three specifications. Consistent with Hall and Liebman (2000), Perry and Zenner estimate lower growth of salaries for affected firms and evidence of an increase in the sensitivity of bonus and total compensation to stock performance after 1993.

Rose and Wolfram (2002) point out that using 1992 levels of pay to classify firms creates statistical correlation between compensation and the variable that indicates whether firms are affected by the regulation that does not correspond to an effect of the change in the law: Firms that had higher compensation levels in 1992 had different characteristics than low compensation firms and these characteristics likely influence their pay growth rate. They propose an alternative estimation using differences in differences that circumvents this problem.[29] To correct for potential bias, Rose and Wolfram construct their indicator for affected firms using predicted instead of observed compensation for years after 1993: They use data in 1992 to construct what the level of compensation of each firm in the sample would have been had there not been a change in regulation. They construct the indicator based on combined salary and bonus payments, since they find that using only salary or broader measures of compensation constitutes a less precise predictor. Based on their classification methodology, the sample mean growth rates of compensation after 1993 were higher for affected firms than for those that were unaffected. An important feature in Rose and Wolfram's analysis is that they include data on the decisions of qualification of salary, bonus, and stock plans (although this is limited to a small sample of firms, based on a consultant firm survey). Stock option plan qualification is the most common choice; in their sample, this implies a tax savings of about 25 percent of ex ante total compensation. About two-thirds of the sample of firms had qualified their stock plans by 1997. Rose and Wolfram find that affected firms were more likely than unaffected firms to qualify their long-term incentive plans (more than twice as likely) and bonus plans (three times more likely), but both groups qualify their option plans in the same proportion. They find evidence of salary compression at the $1 million level after 1993 and a flattening of the distribution of cash payments (salary plus bonus) for firms that choose to qualify bonus plans. However, they do not find strong statistical evidence for an overall compression effect for all firms. Hence, they find no evidence of a perverse consequence of the law suggested in the financial press, whereby the argument was that nonaffected firms raised their salaries to $1 million because the limit was in fact made into

[29] See, also, Rose and Wolfram (2000) for a detailed explanation.

an "acceptable" standard by the new law. Interestingly, Rose and Wolfram find that firms that choose to qualify their option plan have higher growth of salary, bonus, and total compensation after 1993 than both unaffected firms and affected firms that do not qualify. Their estimates are, in general, very noisy, but their evidence is consistent with the hypothesis that qualification and limiting the compensation growth are alternative means of responding to political pressures on executive compensation.

Regulating Corporate Governance

Several regulations passed in the last 15 years have focused on improving corporate governance in U.S. firms. I now summarize the most important initiatives.

In 1992 the SEC increased the disclosure requirements in proxy statements, asking firms to include detailed information about the compensation of the CEO, chief financial officer, and the three other highest-paid executives. Along with those details, the statements had to include an explanation of the compensation policies of the firm, as well as performance measures of the firm in the past five years. In particular, firms were required to report the value of option grants given to the executives. Options could be valued at the time of grant using several alternatives. If Black-Scholes valuation was used, companies were not required to specify the value of the parameters used in the formula, such as the risk-free interest rate, expected volatility, and dividend yield, or any adjustment to the valuation made to take into account the nontransferability of the options and the risk of forfeiture. Alternatively, firms could choose any other accepted pricing methodology, even simply the value of the options under the assumption that stock prices would grow at 5 percent or 10 percent annually during the term of the option (the "potential" or "intrinsic" value of options). In November 1993, the SEC amended its rules and required the disclosure of the parameters used for the valuation and details of any discount.[30] The reforms of 1992 also expanded the set of allowable topics for shareholder proposals to include executive compensation and decreased the minimum share or capital stake necessary to initiate a proxy vote.

Following the accounting fraud scandals at Enron, Tyco, WorldCom, and a number of other firms, the government passed the Sarbanes-Oxley Act in July 2002.[31] The Act had several consequences for corporate governance practices, mainly the requirement of independent accounting auditing, mandatory executive certification of financial reports (accompanied by an increase in penalties for corporate fraud), forfeiture of certain bonuses to executives

[30] Murphy (1996) finds evidence that, in the year in which they had a choice, managers chose the valuation method that minimized the value of their pay.

[31] In November 2001, after weeks of SEC investigations, Enron filed a restatement of its financial results that revealed that the company had underreported its debt exposure.

after a financial restatement resulting from malpractice, or the prohibition of personal loans extended by the firm to the executives. In November 2003, the SEC approved proposals by the NYSE and NASDAQ to guarantee the independence of directors, compensation committees, and auditors.

In 2004, the Financial Accounting Standards Board (FASB) modified their recommendations for the valuation of stock grants: Reporting the fair value of options became the norm, eliminating the previous alternative of reporting their intrinsic value, and expensing requirements were extended to options granted with an exercise price equal to or higher than the market price at the time of granting (FAS 123[R]). In 2006, the SEC made these recommendations compulsory. At the same time, it increased the disclosure requirements of compensation of the executives (salary and bonus payments for the past three years, a detailed separate account of stock and option grants, as well as holdings that originated in previous year's grants, including their vesting status, and retirement and post-employment compensation details). The SEC also demanded an explanation of compensation objectives and implementation in "plain English." It required a classification of the members of the board of directors and of the compensation committees as independent or not independent, with an explicit statement of the arguments used for this classification. Finally, it asked for the disclosure of compensation of directors following rules similar to those for the top executives.

Recently, the proposal of making mandatory a "say on pay" nonbinding vote of shareholders on the compensation of the CEO has received attention both from the media and politicians, and has been voluntarily adopted by a few U.S. corporations.[32] This proposal is in line with the above described efforts of making compensation practices as transparent as possible to shareholders.

Quantifying Effects of Regulation on Corporate Governance

Most of the above described changes in regulation could directly or indirectly influence compensation contracts, as well as firm performance. A few studies have tried to quantify these potential effects.

Johnson, Nelson, and Shackell (2001) document that the 1992 SEC reforms related to shareholder participation translated into an increase in the number of shareholder-initiated proposals on CEO pay. However, they do not conclude that the probability of a proposal is higher for firms with poor incentive alignment.[33] Instead, this probability of proposal is higher for firms with higher levels (or lower sensitivity) of CEO compensation. They attribute this correlation to the higher "exposure" of the compensation practices of firms

[32] The "say on pay" practice is already in place in the United Kingdom, the Netherlands, and Australia. See, for example, the article "Fair or Foul" (*The Economist* 2008a).

[33] Poor incentive alignment is approximated by the proportion of compensation that is not predictable by observable variables, following Core, Holthausen, and Larcker (1999).

with higher salaries. They find, however, a positive effect on corporate governance when they look at the probability of acceptance of the proposals: it is higher in firms with poor incentive alignment and higher for institutional investor proposals.

Chhaochharia and Grinstein (2007) provide some evidence on the effects of SOX and of the changes in the stock exchange independence requirements, as evaluated by investors in the stock market. They analyze the effects of the announcements of those regulatory changes on the stock prices of firms with different compliance levels. They construct portfolios of firms based on their degrees of compliance and find that less compliant firms (for example, firms that restated their financial statements or that do not have independent boards) earn abnormal returns of about 6 to 20 percent around the announcement of the new rules. They also report evidence that the market believes that small firms find it more difficult to comply with requirements for internal control and independence of directors: Their portfolios have small negative abnormal returns after the announcements, in contrast with positive abnormal returns of larger noncompliant firms.

The imposition of the $1 million limit with the passing of section 162(m) discussed earlier arguably could have consequences for corporate governance—on top of any direct effect on executive compensation. The evidence discussed in the previous subsection points to an increase in sensitivity of pay, which may have translated to better alignment of the incentives of the CEO with those of shareholders. This hypothesis is consistent with the evidence in Maisondieu-Laforge, Kim, and Kim (2007): They find that stock returns and operating returns improve for firms affected by the cap. However, one may argue that compensation committees may have been compelled to forego some of the subjective components in the granting of bonuses and options in order to qualify compensation plans for tax exemptions. These committees may have been compelled to qualify their compensation plans solely for tax savings purposes, or to comply with what may have been viewed as the "correct" practice after section 162(m) was approved. Hayes and Schaefer (2000) find that the part of CEO compensation that is unexplained by observable performance measures is a good predictor for future performance. This suggests that discretion of the compensation board may in fact be rewarding good executive performance, and the requirements for qualification of bonus schemes may be distorting good compensation practices.

5. CONCLUSION

Real life compensation contracts are complicated. Even in the absence of illegal actions on the part of CEOs, it is easy to find examples in which the incentives of managers are not perfectly aligned with those of the shareholders. The recent cases of fraudulent behavior, such as the accounting scandals

at Enron, the backdating of options, or insider trading, obviously demand government intervention, as does any other breach of the law. However, a close look at the problem of CEO pay design demonstrates that providing incentives is a complicated matter and that many seemingly unintuitive features of compensation packages may, in fact, be helpful in solving incentive problems. There is ample evidence that, despite rising levels of CEO pay in the last two decades, improved corporate governance practices have translated into stronger ties between pay and performance over the same period. Recent regulation of corporate governance seems to have been aiding market-driven changes in internal control. However, it is also easy to find indications that some of the compensation regulation may impose unnecessary burdens and distortions on firms.

REFERENCES

Acharya, Viral V., Kose John, and Rangarajan K. Sundaram. 2000. "Contract Renegotiation and the Optimality of Resetting Executive Stock Options." *Journal of Financial Economics* 57: 65–101.

Associated Press. 2008. "CEO Pay Keeps Rising Even as Economy Slows." *Richmond Times Dispatch*. http://www.inrich.com/cva/ric/search.apx.-content-articles-RTD-2008-06-16-0104.html [16 June].

Baker, George P., and Brian J. Hall. 2004. "CEO Incentives and Firm Size." *Journal of Labor Economics* 22 (October): 767–98.

Bebchuk, Lucian, and Jesse Fried. 2003. "Executive Compensation as an Agency Problem." *Journal of Economic Perspectives* 17 (January): 71–92.

Bebchuk, Lucian, and Jesse Fried. 2004. *Pay without Performance: The Unfulfilled Promise of Executive Compensation*. Cambridge, Mass.: Harvard University Press.

Beck, Rachel, and Matthew Fordahl. 2008a. "CEO Pay Chugs Up in '07 Despite Economy." http://www.hermes-press.com/CEO_pay1.htm [16 June].

Beck, Rachel, and Matthew Fordahl. 2008b. "CEO Pay Climbs Despite Companies' Struggles." *USA Today*. http://www.usatoday.com/money/companies/management/2008-06-15-ceo-pay_N.htm [20 June].

Bennedsen, Morten, Francisco Pérez-González, and Daniel Wolfenzon. 2007. "Do CEOs Matter?" Copenhagen Business School Working Paper.

Brenner, Menachem, Rangarajan Sundaram, and David Yermack. 2000. "Altering the Terms of Executive Stock Options." *Journal of Financial Economics* 57 (July): 103–28

Bolton, Patrick, José Scheinkman, and Wei Xiong. 2006. "Executive Compensation and Short-termist Behavior in Speculative Markets." *Review of Economic Studies* 73: 577–610.

Carter, Mary Ellen, and Luann J. Lynch. 2001. "An Examination of Executive Stock Option Repricing." *Journal of Financial Economics* 61 (August): 207–25.

Chance, Don, Raman Kumar, and Rebecca Todd. 2000. "The 'Repricing' of Executive Stock Options." *Journal of Financial Economics* 57: 129–54.

Chhaochharia, Vidhi, and Yaniv Grinstein. 2007. "Corporate Governance and Firm Value: The Impact of the 2002 Governance Rules." *Journal of Finance* 62: 1789–825.

Celentani, Marco, and Rosa Loveira. 2006. "A Simple Explanation of the Relative Performance Evaluation Puzzle." *Review of Economic Dynamics* 9: 525–40.

Clementi, Gian Luca, Thomas F. Cooley, and Cheng Wang. 2006. "Stock Grants as a Commitment Device." *Journal of Economic Dynamics and Control* 30 (November): 2191–216.

Core, John E., Robert W. Holthausen, and David F. Larcker. 1999. "Corporate Governance, Chief Executive Officer Compensation, and Firm Performance." *Journal of Financial Economics* 51: 371–406.

Cuñat, Vicente, and Maria Guadalupe. 2004. "Executive Compensation and Competition in the Banking and Financial Sectors." Working Paper (April).

Cuñat, Vicente, and Maria Guadalupe. 2005. "How Does Product Market Competition Shape Incentive Contracts?" *Journal of the European Economic Association* 3: 1058–82.

Cuñat, Vicente, and Maria Guadalupe. 2006. "Globalization and the Provision of Incentives Inside the Firm." Columbia University Working Paper (November).

Edmans, Alex, Xavier Gabaix, and Augustin Landier. Forthcoming. "A Multiplicative Model of Optimal CEO Incentives in Market Equilibrium." *Review of Financial Studies*.

Frydman, Carola. 2005. "Rising Through the Ranks: The Evolution of the Market for Corporate Executives, 1936–2003." Working Paper (November).

Frydman, Carola, and Raven Saks. 2007. "Historical Trends in Executive Compensation, 1936–2003." Harvard University Working Paper (November).

Gabaix, Xavier, and Augustin Landier. 2008. "Why Has CEO Pay Increased So Much?" *The Quarterly Journal of Economics* 123: 49–100.

Garen, John. 1994. "Executive Compensation and Principal-Agent Theory." *Journal of Political Economy* 102 (December): 1175–99.

Gibbons, Robert, and Kevin J. Murphy. 1990. "Relative Performance Evaluation for Chief Executive Officers." *Industrial and Labor Relations Review* 43: 30S–51S.

Gibbons, Robert, and Kevin J. Murphy. 1992. "Optimal Incentive Contracts in the Presence of Career Concerns: Theory and Evidence." *Journal of Political Economy* 100 (June): 468–505.

Gompers, Paul A., and Andrew Metrick. 2001. "Institutional Investors and Equity Prices." *Quarterly Journal of Economics* CXIV: 229–60.

Gompers, Paul A., Joy L. Ishii, and Andrew Metrick. 2003. "Corporate Governance and Equity Prices." *Quarterly Journal of Economics* 118 (February): 107–55.

Grossman, S., and O. Hart. 1983. "An Analysis of the Principal-Agent Problem." *Econometrica* 51: 7–45.

Hall, Brian, and Jeffrey Liebman. 1998. "Are CEOs Really Paid Like Bureaucrats?" *Quarterly Journal of Economics* 113 (August): 653–91.

Hall, Brian, and Jeffrey Liebman. 2000. "The Taxation of Executive Compensation." Working Paper 7596. Cambridge, Mass.: National Bureau of Economic Research. (March).

Hall, Brian J., and Kevin J. Murphy. 2000. "Optimal Exercise Prices for Executive Stock Options." *American Economics Review Papers and Proceedings* 90: 209–14.

Haubrich, Joseph. 1994. "Risk Aversion, Performance Pay, and the Principal-Agent Problem." *Journal of Political Economy* 102 (April): 258–76.

Hayes, Rachel, and Scott Schaefer. 2000. "Implicit Contracts and the Explanatory Power of Top Executive Compensation for Future Performance." *RAND Journal of Economics* 31: 273–93.

Hermalin, Benjamin E., and Michael S. Weisbach. 1998. "Endogenously Chosen Boards of Directors and Their Monitoring of the CEO." *American Economic Review* 88 (April): 96–118.

Holmström, Bengt. 1999. "Managerial Incentive Problems: A Dynamic Perspective." *Review of Economic Studies* 66 (January): 169–82.

Holmström, Bengt, and Steven Kaplan. 2003. "The State of U.S. Corporate Governance: What's Right and What's Wrong?" *Journal of Applied Corporate Finance* 15: 8–20.

Jensen, Michael, and Kevin J. Murphy. 1990a. "Performance Pay and Top-Management Incentives." *Journal of Political Economy* 98 (April): 225–64.

Jensen, Michael C., and Kevin J. Murphy. 1990b. "CEO Incentives: It's Not How Much You Pay, But How." *Harvard Business Review* 3 (May-June).

Jensen, Michael, Kevin J. Murphy, and Eric Wruck. 2004. "Remuneration: Where We've Been, How We Got to Here, What are the Problems, and How to Fix Them." Harvard NOM Working Paper 04-28. (July).

Jenter, Dirk, and Fadi Kanaan. 2006. "CEO Turnover and Relative Performance Evaluation." Working Paper 12068. Cambridge, Mass.: National Bureau of Economic Research. (March).

Johnson, Marilyn, Karen Nelson, and Margaret Shackell. 2001. "An Empirical Analysis of the SEC's 1992 Proxy Reforms on Executive Compensation." Stanford University Graduate School of Business Research Paper 1679.

Kadan, Ohad, and Jeroen Swinkels. 2008. "Stocks or Options? Moral Hazard, Firm Viability, and the Design of Compensation." *Review of Financial Studies* 21 (January): 451–82.

Kaplan, Steven N., and Bernadette A. Minton. 2006. "How has CEO Turnover Changed? Increasingly Performance Sensitive Boards and Increasingly Uneasy CEOs." Working Paper 12465. Cambridge, Mass.: National Bureau of Economic Research.

Kole, Stacey R. 1997. "The Complexity of Compensation Contracts." *Journal of Financial Economics* 43 (January): 79–104.

Kuhnen, Camelia, and Jeffrey Zwiebel. 2008. "Executive Pay, Hidden Compensation and Managerial Entrenchment." Working Paper (July).

Lucas, Robert E., Jr. 1978. "On the Size Distribution of Business Firms." *Bell Journal of Economics* 9: 508–23.

Margiotta, Mary M., and Robert A. Miller. 2000. "Managerial Compensation and the Cost of Moral Hazard." *International Economic Review* 41 (August): 669–719.

Maisondieu-Laforge, Olivier, Yong H. Kim, and Young Sang Kim. 2007. "Financial Contracting and Operating Performance: The Case for OBRA and Efficient Contracting." *Corporate Ownership & Control* 4: 217–27.

Miller, Merton H., and Myron S. Scholes. 2002. "Executive Compensation, Taxes, and Incentives." In *Selected Works of Merton H. Miller: A Celebration of Markets*, edited by Bruce D. Grundy. Chicago: The University of Chicago Press, 254–78.

Murphy, Kevin J. 1996. "Reporting Choice and the 1992 Proxy Disclosure Rules." *Journal of Accounting, Auditing, and Finance* 11: 497–515.

Murphy, Kevin J. 1999. "Executive Compensation." In *Handbook of Labor Economics*, edited by Orley Ashenfelter and David Card. New York: Elsevier Science North Holland, 2485–563.

Perry, Tod, and Marc Zenner. 2001. "Pay for Performance? Government Regulation and the Structure of Compensation Contracts." *Journal of Financial Economics* 62 (December): 453–88.

Prescott, Edward S. 2003. "Firms, Assignments, and Earnings." Federal Reserve Bank of Richmond *Economic Quarterly* 89 (Fall): 69–81.

Rose, Nancy, and Catherine Wolfram. 2000. "Has the 'Million-Dollar Cap' Affected CEO Pay?" *American Economic Review Papers and Proceedings* 90 (May): 197–202.

Rose, Nancy, and Catherine Wolfram. 2002. "Regulating Executive Pay: Using the Tax Code to Influence Chief Executive Officer Compensation." *Journal of Labor Economics* 20 (April): S138–S175.

Rosen, Sherwin. 1981. "The Economics of Superstars." *American Economic Review* 71: 845–58.

Rosen, Sherwin. 1982. "Authority, Control and the Distribution of Earnings." *Bell Journal of Economics* 13: 311–23.

Rosen, Sherwin. 1992. "Contracts and the Market for Executives." In *Contract Economics*, edited by Lars Werin and Hans Wijkander. Cambridge, Mass.: Blackwell, 181–211.

Schaefer, Scott. 1998. "The Dependence of Pay-Performance Sensitivity on the Size of the Firm." *The Review of Economics and Statistics* 80 (August): 436–43.

Terviö, Marko. 2008. "The Difference that CEOs Make: An Assignment Model Approach." *American Economic Review* 98 (June): 642–68.

The Economist. 2008a. "Fair or Foul." http://www.economist.com/business/displaystory.cfm?story_id=11543754 [12 June].

The Economist. 2008b. "Pay Attention." http://www.economist.com/business/displaystory.cfm?story_id=11543665 [12 June].

United States Senate, Joint Committee on Taxation. 2006. "Present Law and Background Relating to Executive Compensation." Staff report JCX-39-06 (5 September).

Wall Street Journal. 2008. "Rising Pay and Unintended Consequences." http://online.wsj.com/article/SB120793959097608465.html?mod=2_1565_leftbox [14 April].

Wang, Cheng. 1997. "Incentives, CEO Compensation, and Shareholder Wealth in a Dynamic Agency Model." *Journal of Economic Theory* 76 (September): 72–105.

Weinberg, John A. 2003. "Accounting for Corporate Behavior." Federal Reserve Bank of Richmond *Economic Quarterly* 89 (Summer): 1–20.