Limited Participation and the Neutrality of Money

Stephen D. Williamson

Money is useful in overcoming two types of frictions. First, in barter exchange, money helps to mitigate double-coincidence frictions that arise in developed economies where economic agents are specialized in production and consumption. Second, in trades involving credit, information frictions may imply that one economic agent has difficulty getting another economic agent to accept his or her IOUs in exchange for goods and services. In an economy with monetary exchange, much more can be achieved than in an economy without money. Even so, one of the key lessons of monetary economics is that circumstances exist in which changing the quantity of money will not matter at all for what can be produced and consumed in an economy. For example, governments sometimes engage in currency reforms, particularly in circumstances where there has been a recent history of high inflation.

One such instance occurred in January 2005, when Turkey introduced a new Turkish lira, equivalent in all respects to the old Turkish lira, except that one new lira trades for one million old lira. That is, the central bank of Turkey declared itself willing to exchange old lira for new lira at a rate of one million to one. What would be the effects of this? To help frame the problem, suppose that the U.S. government were to announce a currency reform, where a new U.S. dollar was introduced, defined as being equivalent to 10 old U.S. dollars. Suppose also that the Federal Reserve did not otherwise change its behavior. The result would be that the money stock and all prices and wages in terms of the new U.S. dollar would be one-tenth of what they would have been under the old U.S. dollar, and all real economic activity would be unchanged.

Chester A. Phillips Professor of Financial Economics, University of Iowa, and Visiting Scholar, Federal Reserve Bank of Richmond. Helpful comments and suggestions from Andreas Hornstein, Jennifer Sparger, Alex Wolman, and Roy Webb are gratefully acknowledged. The views here are the author's and should not be attributed to the Federal Reserve Bank of Richmond or the Federal Reserve System.

Money would clearly be neutral under these circumstances. That is, changing the quantity of money by simply redefining the units in which we measure the money stock can have no implications for real economic activity.

However, when changing the stock of money through an open market operation, the Federal Reserve System is hardly carrying out a currency reform. For example, when the Fed conducts an open market operation, the economic agents on the receiving end of this transaction typically are large financial institutions that are not directly connected to all other economic agents in the economy through exchange. Initially, an open market operation can affect only the financially interconnected sector of the economy—mainly banks and other financial intermediaries and the economic agents who transact frequently with these institutions. In contrast to what happens in a currency reform, a typical open market operation will, in the short run, have different effects in the financially interconnected sector of the economy from what happens in the decentralized sector of the economy. This difference will be important for short-run movements in interest rates, aggregate output, and the distribution of wealth across the population.

The idea that monetary policy matters in the short run because of financial disconnectedness in the economy is captured in *limited participation models.* The first models of this type were constructed by Grossman and Weiss (1983) and Rotemberg (1984). These early contributions were heterogeneous-agent models that proved to be very difficult to work with due to complications in tracking the distribution of wealth across the population over time. Lucas (1990) finessed these complications by working with a representative-household construct, as did Fuerst (1992). Later important contributions were made by Alvarez and Atkeson (1997) and Alvarez, Atkeson, and Kehoe (2002). Much of this research, which focuses mainly on the asset pricing implications of limited participation, is weak on the microfoundations of money and, as a result, may be misleading. More recently, Williamson (forthcoming) and Shi (2004) constructed monetary search models that treat monetary exchange seriously and permit the study of the role of limited participation in generating short-run nonneutralities of money.

This article reviews the literature on limited participation and points out new directions for research by constructing and building on a baseline limited participation model. The baseline model is a cash-in-advance model similar to the one studied in Lucas (1990).[1] I first show how limited participation provides an explanation for the liquidity effect—the short-run negative response of the nominal interest rate to an open market purchase. The baseline model does not provide a rationale for monetary policy though, as in this setup monetary policy has no implications for real economic variables and economic

[1] I do not provide a rigorous microfoundation for monetary exchange in this model, but it would be straightforward to do this, for example, along the lines of Williamson (2004a).

welfare. As I discuss in Section 2, an extension of the baseline model along the lines of Fuerst produces a short-run nonneutrality of money in that an open market purchase of interest-bearing government securities by the central bank reduces the nominal interest rate, increases the real wage, and increases real output. In that environment, however, it would be optimal for the monetary authority not to intervene in the economy.

The conclusion from examining the implications of the first two versions of the model is that, while these variations provide an explanation for the liquidity effect, they do not teach us much about the real effects of monetary policy or how to conduct policy.

In Section 3, in the third incarnation of the model, I extend the framework in a new direction by permitting a persistent distributional effect of monetary policy. Here, money is nonneutral whether monetary intervention is anticipated or not (a feature not shared with the previous two incarnations of this model). In this version of the model, characterizing the effects of monetary policy is difficult without getting outside the scope of this article, but determining the model's implications for optimal monetary policy is relatively straightforward. Here, optimal monetary policy is in one sense a Friedman rule (the nominal interest rate is always zero at the optimum), but in another sense is much more complicated than a typical Friedman rule. This complication arises because the goal of the monetary authority is to control monetary conditions in the financially disconnected sector of the economy, but monetary control can be achieved only indirectly—through intervention in the financially connected sector of the economy. Finally, Section 4 serves as a conclusion.

1. LIMITED PARTICIPATION AND THE LIQUIDITY EFFECT

This section contains the baseline model—closely related to the model in Lucas (1990)—used throughout this article. Lucas considered a simple asset-pricing model without production, while my model allows for production and endogenous labor supply. As well, there are some minor differences from Lucas's work in how I specify asset markets.

The Representative Household

In the model, a representative infinitely lived household maximizes

$$E_0 \sum_{t=0}^{\infty} \beta^t [u(c_t) - v(n_t)],$$

where E_0 is the expectation operator conditional on information in period 0; β is the household's discount factor with $0 < \beta < 1$; c_t is the household's consumption; and n_t is household labor supply. Assume that $u(\cdot)$ is strictly increasing, strictly concave, and twice differentiable with $u'(0) = \infty$, and that $v(\cdot)$ is strictly increasing, strictly convex, and twice differentiable with $v'(0) = 0$ and $v'(h) = \infty$ where h is the household's endowment of time each period. The household has a technology that permits it to produce one unit of the perishable consumption good in period t for each unit of labor supplied.

One of the innovations in Lucas (1990) was to model a household as having many agents, with household members engaged in different activities during each period. This device is used here, and its purpose is to make the model analytically tractable in that monetary policy in this model will only cause changes in the distribution of wealth within the household and within a period, not persistent changes in the distribution of wealth across economic agents. Thus, in the baseline version of the model, the household consists of three agents: a worker, a shopper, and a financial transactor.

As is typical in models with cash-in-advance constraints, the timing of transactions within a period is critical to how the model works. At the beginning of the period, the household has M_t units of money on hand and must then decide how to split these money balances between the shopper, who will go to the goods market to purchase consumption goods from other households, and the financial transactor, who will go to the asset market to purchase assets. Let X_t denote the quantity of money that the household sends to the goods market with the shopper, where $X_t \leq M_t$. For the shopper, the value of goods purchased cannot exceed X_t; that is, the shopper faces the cash-in-advance constraint

$$P_t c_t \leq X_t, \tag{1}$$

where P_t is the price level, the price of consumption goods in terms of money. Now, in the asset market I will assume that there is only one asset bought and sold, which is a nominal bond issued by the government. One nominal government bond issued in period t sells at a price q_t in terms of money and is a promise to pay one unit of money at the end of the period. Bonds must be purchased with money, so the financial transactor, like the shopper, faces a cash-in-advance constraint, which in this case is

$$q_t B_t \leq M_t - X_t. \tag{2}$$

The worker stays at the household's location where he produces and sells goods to other households. As is usual in cash-in-advance models, I assume that the household cannot consume its own output and that money acquired by the household from the sale of its output cannot be used within the period to purchase consumption goods or government bonds. The household then faces the budget constraint

$$P_t c_t + M_{t+1} + q_t B_t \leq M_t + B_t + P_t z_t n_t - \Upsilon_t, \tag{3}$$

where M_{t+1} is the quantity of money that the household carries into period $t + 1$ and Υ_t is a lump-sum tax that the household pays in money to the government at the end of the period. The left-hand side of the budget constraint (3) consists of the value of consumption goods purchased by the shopper, plus money balances at the end of the period, plus the value of bonds purchased by the financial transactor. On the right-hand side is the quantity of money possessed by the household at the beginning of the period, plus the total payoff on government bonds held by the household, plus the proceeds from sales of goods by the worker, minus the lump-sum tax paid to the government. It may seem unusual to have government bond purchases B_t appear on the right-hand and left-hand sides of the budget constraint (3). However, these are within-period bonds for which the household gives up $q_t B_t$ units of money on the asset market and receives B_t units of money as a payoff at the end of the period.

The Government

Each period, the government must choose the quantity of nominal bonds to issue, which I denote by $\overline{B}_t$ (I will use $\overline{\textit{overbar}}$ throughout to denote the supplies of assets determined by the government). I will assume that

$$\overline{B}_t = \theta_{t+1}\overline{M}_t, \tag{4}$$

where $\overline{M}_t$ is the quantity of money outstanding at the beginning of period t and θ_{t+1} is a random variable that is not realized until after the shopper and financial transactor have left the household in period t. At this point, perceptive readers might quarrel with the assumption that the government behaves randomly. This assumption proves useful in making my argument, and I will comment later on what happens if θ_{t+1} is a choice variable for the central bank.

To obtain a clean policy experiment in the model, I will assume that the government sets the lump-sum tax Υ_t so that the money stock at the end of the period is identical to the beginning-of-period money stock. That is,

$$\Upsilon_t = (1 - q_t)\overline{B}_t. \tag{5}$$

Given equation (5), the money stock will remain fixed for all time, and θ_t will not affect the money growth rate. The following is one interpretation of how policy is conducted in this model, consistent with the notion that it is desirable here to set up the policy experiment so that it captures monetary policy and is not some mix of fiscal and monetary policies. Each period, the fiscal authority issues $\hat{\theta}\overline{M}_t$ nominal bonds, where $\hat{\theta}$ denotes the maximum possible realization of θ_{t+1}. Then, the central bank determines (randomly) how much of the bond issue to acquire from the fiscal authority, purchasing $\left(\hat{\theta} - \theta_{t+1}\right)\overline{M}_t$ bonds, thus leaving $\theta_{t+1}\overline{M}_t$ bonds to be purchased by the public at price q_t. Then, at the end of the period, the fiscal authority has $q_t\theta_{t+1}\overline{M}_t$ units of money acquired from bond sales and must pay the bondholders $\theta_{t+1}\overline{M}_t$ units of money, as promised. It then makes up the difference $(1-q_t)\theta_{t+1}\overline{M}_t$ through a lump-sum tax on the representative household, so that the tax is given by (5). Transactions between the fiscal authority and the central bank merely yield accounting entries, and the central bank's account balance with the fiscal authority is reset to zero at the end of each period.

Assume that information is not transmitted during the period between the asset market and the goods market, so workers and shoppers do not learn θ_{t+1} until the end of the period after all decisions have been made.

Optimization and Equilibrium

To specify the household's optimization problem in a tractable way, it proves useful to divide the left-hand and right-hand sides of equations (1) through (3) by $\overline{M}_t$ and let lower-case variables denote the corresponding upper-case variable scaled by the money supply, $p_t \equiv \frac{P_t}{\overline{M}_t}$, for example. For convenience, assume for now that θ_t is an i.i.d. random variable. Further, drop t subscripts and let primes denote variables dated $t+1$. Then, I can specify the representative household's optimization problem as a dynamic program, where $V(m, \theta)$ is the household's value function. The household solves

$$V(m,\theta) = \max_{x,c,n} \left[u(c) - v(n) + \beta E_\theta \max_{b,m'} E_{\theta'} V(m', \theta') \right] \tag{6}$$

subject to

$$pc \leq x, \tag{7}$$

$$qb \leq m - x, \text{ and} \tag{8}$$

$$pc + m' + qb \leq m + b + pn - \tau. \tag{9}$$

In the objective function (6), E_θ is the expectation operator conditional on information before θ' is known, while $E_{\theta'}$ conditions on θ'.

To solve the household's problem, first note that the optimal choice of b in the inner maximization problem in (6) gives

$$b = \frac{m-x}{q}, \text{ if } q \leq 1, \text{ and} \tag{10}$$

$$b = 0, \text{ if } q > 1.$$

Given that the government always issues a strictly positive quantity of nominal bonds, we must have $q \leq 1$ in equilibrium, so confining attention to this case and substituting for b in (9) using (10) gives

$$pc + m' \leq m + \frac{(m-x)(1-q)}{q} + pn - \tau. \tag{11}$$

I can then specify the household's problem as solving (6) subject to (7) and (11). Let λ_1 and λ_2 denote the multipliers associated with constraints (7) and (11), respectively. From the choice of m' in the inner maximization in (6), I obtain the first order condition (assuming the value function is strictly concave and differentiable, and using the relevant envelope condition),

$$-\lambda_2 + \beta E_{\theta'}\left(\frac{\lambda_2'}{q'}\right) = 0, \tag{12}$$

and the choices of x, c, and n in the outer maximization problem in (6) give the following first order conditions, respectively:

$$\lambda_1 - E_\theta\left[\lambda_2\left(\frac{1}{q} - 1\right)\right] = 0, \tag{13}$$

$$u'(c) - p(\lambda_1 + E_\theta\lambda_2) = 0, \text{ and} \tag{14}$$

$$-v'(n) + pE_\theta\lambda_2 = 0. \tag{15}$$

In equilibrium, the bond market clears, or

$$b = \theta'; \tag{16}$$

the representative household willingly holds the existing stock of money, or

$$m = 1; \tag{17}$$

and the market for consumption goods clears, or

$$c = n. \tag{18}$$

Given the assumption that θ is an i.i.d. random variable, I can solve for an equilibrium in which x, c, n, and p are constant. First assume that the cash-in-advance constraint (7) binds. Then, (12) through (18) give

$$x = 1 - \beta E(\theta), \tag{19}$$

$$v'(c) - \beta u'(c) = 0, \tag{20}$$

$$n = c, \tag{21}$$

$$p = \frac{1 - \beta E(\theta)}{c}, \text{ and} \tag{22}$$

$$q = \frac{\beta E(\theta)}{\theta}. \tag{23}$$

In (19) through (23), $E(\theta)$ is the expected value of θ. Note, from (19) that the fraction of money balances allocated to the shopper for the purchase of consumption goods is decreasing in the expected size of the government's open market operation. Monetary policy has no effect on any variables of consequence, as equation (20) determines consumption (equal to labor supply) and, thus, c is independent of θ and the distribution of θ. The only effect of θ is on the price of the nominal bond q in equation (23). Clearly q is decreasing in θ, so that the nominal interest rate increases as θ increases. This is the liquidity effect—if the government withdraws more outside money through an open market sale, the nominal interest rate will be higher.

An interesting feature of the setup here is that I have designed the policy experiment to imply no Fisher effect—the positive effect of money growth and inflation on the nominal interest rate. Because monetary policy leaves the money supply constant over time, the only effect on the nominal interest rate is the liquidity effect. Note that the presence of $E(\theta)$ in equations (19), (22), and (23) has nothing to do with the Fisher effect. Instead, if $E(\theta)$ is high, then it is expected that a higher quantity of bonds will be sold to private agents, and the household therefore also predicts that the expected payoff from holding government bonds will be higher. Thus, the household will tend to allocate more cash to the asset market as opposed to the goods market (x declines in equation (19)). Both the price level and the price of the nominal bond are in turn determined in part by x, as (22) and (23) indicate.

It is straightforward to show that, given the solutions (19) through (23), we will have $\lambda_1 > 0$, so that the cash-in-advance constraint binds. As well, my solution requires that $q \leq 1$ in equilibrium, so from (23) we require that $\theta \geq \beta E(\theta)$ for all realizations of the random variable θ.

The implication of this model is that, while monetary policy can produce variability in the nominal interest rate, policy is irrelevant for economic welfare as it does not affect consumption and employment. The fact that the asset market and goods market are segmented implies that nominal interest rate movements will have no real effects. Note as well that the price level is in some sense "sticky," as in equation (22) p does not depend on θ—monetary policy can change the supply of liquidity in the asset market but has no effect on the quantity of money in the goods market. However, p depends on $E(\theta)$, so anticipated monetary policy matters for the determination of the price level, though not for any real variables of consequence.

If the model as it stands were a good description of reality, we would conclude that central bank intervention in asset markets was a useless exercise. The central bank might just as well do nothing rather than cause the nominal interest rate to fluctuate, but doing so certainly does not cause any harm. The liquidity effect on nominal interest rates is an important element in the religion of central bankers [2], and Lucas (1990) provides an explanation for the liquidity effect. However, the model does nothing to show why monetary policy matters

[2] In my opinion, belief in the liquidity effect by central bankers is much like religious belief: It seems impossible to disprove its existence because of the endogeneity of the money stock brought about by endogenous policy and the endogenous response of the banking system to exogenous shocks to the economy. If monetary policy were exogenous, then it would be straightforward to measure the effect of a money shock on nominal interest rates. The problem is that the right natural experiment does not appear to have occurred in practice.

or how it should be conducted. To address these matters, I need to modify the baseline model in ways that make money nonneutral.

2. LIMITED PARTICIPATION WITH REAL EFFECTS ON OUTPUT

In this section, I alter the baseline model constructed in the previous section to produce nonneutralities of short-run monetary policy. The basic idea comes from Fuerst (1992).

In the baseline model, I assumed that a household must buy consumption goods from other households with cash acquired in advance of the current period. The extension here is to assume that the household can produce only with labor supplied by other households and that labor, like consumption goods, is subject to a cash-in-advance constraint. To be more explicit, the timing works as follows: At the beginning of the period, the representative household has M_t units of money and sends X_t units of money with the shopper to the goods market. The worker also leaves the household at the same time as the shopper to sell labor to other households. The financial transactor is left behind to trade on the asset market and to purchase y_t units of labor at the real wage rate of w_t from other households, using the remaining quantity of money balances, $M_t - X_t$. Using the same notation as in the previous section, replace the constraint (8) with

$$qb + pwy \leq m - x, \tag{24}$$

and rewrite the household's budget constraint as

$$pc + m' + qb + pwy \leq m + b + pwn + py - \tau. \tag{25}$$

The representative household's dynamic programming problem is now

$$V(m, \theta) = \max_{x} E_{\theta} \left\{ \max_{c,n,y,b,m'} \left[u(c) - v(n) + \beta V(m', \theta') \right] \right\}, \tag{26}$$

subject to (7), (24), and (25). Note that on the right-hand side of the Bellman equation (26) the household must choose x, the fraction of cash sent with the shopper, before the central bank intervenes in the asset market, but all other household choices are made with knowledge of the central bank's action θ', albeit with x already locked in place. That is, θ' is revealed to the shopper through the price p and to the worker through the real wage rate w.

Now, given that θ is an i.i.d. random variable and that the central bank sets the lump sum tax at the end of the period according to (5) so as to keep the aggregate money stock constant from period to period, it is straightforward to characterize the effects of monetary policy on prices, the nominal interest rate, and output. Let λ_1, λ_2, and λ_3 denote the multipliers associated with the constraints (7), (24), and (25), respectively. An equilibrium is the solution to:

$$E_{\theta}(\lambda_1) = E_{\theta}(\lambda_2), \tag{27}$$

$$u'(c) - p(\lambda_1 + \lambda_3) = 0, \tag{28}$$

$$-v'(n) + pw\lambda_3 = 0, \tag{29}$$

$$-w(\lambda_2 + \lambda_3) + \lambda_3 = 0, \tag{30}$$

$$-q\lambda_2 + (1 - q)\lambda_3 = 0, \text{ and} \tag{31}$$

$$-\lambda_3 + \beta E_{\theta'}(\lambda_2' + \lambda_3') = 0, \tag{32}$$

in addition to (7), (24), and (25). This solution is derived from the first order conditions characterizing a solution to the constrained optimization problem on the right-hand side of (26), the relevant envelope conditions, arbitrage conditions, (16) through (18), and the equilibrium condition $y = n$.

From (30) and (31), an interesting feature of the equilibrium is that $w = q$, so that the real wage and the price of the nominal bond are identical, and this equality arises for the following reason. On the one hand, if the financial transactor purchases labor, he gives up money mid-period and receives money in exchange for output at the end of the period. On the other hand, the financial transactor could give up money to purchase a bond with the return on the bond received at the end of the period. In equilibrium, the financial transactor must be indifferent between purchasing labor and acquiring a bond, which requires that $w = q$. Then, if the second cash-in-advance constraint (30) binds so that $\lambda_2 > 0$, we will have $q = w < 1$, or the nominal interest rate is positive and the real wage is less than labor's marginal product. That is, cash received for output cannot be spent until the next period, so no missed profit opportunity necessarily results if the market wage is less than labor's marginal product ($w < 1$).

Now, for convenience, consider an equilibrium where both of the cash-in-advance constraints (7) and (24) bind for all θ. First, given that θ is an i.i.d. random variable, x will be independent of θ in equilibrium. Then, given x, from (7), (24), (25), and (27) through (32), w, q, and c are the solutions to

$$w = q = \frac{1 - x}{x + \theta} \text{ and} \tag{33}$$

$$cv'(c) = \left(\frac{1 - x}{x + \theta}\right)\psi, \tag{34}$$

where ψ is a constant, and, given x and c, the price level is determined by

$$p = \frac{x}{c}. \tag{35}$$

From (33), the real wage and the price of the nominal bond are decreasing in θ. Therefore, a larger open market sale implies a higher nominal interest rate and a lower real wage because of the tightening of the second cash-in-advance constraint (24). Thus, as in the baseline model in the previous section, a liquidity effect exists, but here this effect extends to a change in the wage rate. In addition, a real effect of monetary policy now exists. A smaller open market purchase (smaller θ) relaxes the second cash-in-advance constraint (24) from equation (34), implying that the demand for labor rises, and in equilibrium, the increased demand leads not only to an increase in the wage rate but also to an increase in employment, output, and consumption. That is, in equation (34), the left-hand side is increasing in c, and the right-hand side is decreasing in θ, so that c is decreasing in θ.

The nonneutrality of money here works in an essentially identical manner to the mechanism in Fuerst (1992), which was later adapted in work by Christiano and Eichenbaum (1995). In these models an additional embellishment makes the model seem more plausible. Rather than having the representative household purchase labor directly subject to a cash-in-advance constraint, as is the case here, Fuerst, for example, supposes that a financial intermediary takes cash deposits from the household and makes cash loans to firms and that the firm then pays workers in cash. This construct amounts to the same thing as specified here—labor is purchased subject to a cash-in-advance constraint, highlighting a key defect in this attempt to understand the short-run role for monetary policy. In the United States, few workers are paid in cash, and even if they are, it seems difficult to argue that firms subject to cash-in-advance constraints account for a significant fraction of U.S. employment. Most firms have sufficient access to banking services and financial markets so that they will not face serious cash constraints in paying their workers. To see why this fact is important in the model, suppose that the representative household can issue IOUs in order to pay workers and purchase government bonds on the asset market, with the IOUs being repaid at the end of the period (equivalent in the Fuerst [1992] model to allowing the "bank" to issue within-period IOUs). Then the nominal interest rate is zero in equilibrium, the liquidity effect goes away, and output is constant for all θ in my model.

Another problem with this extension of the baseline model is that it does not provide a rationale for short-run central bank intervention. In spite of the fact that the central bank can cause the nominal interest rate, employment, output, and consumption to fluctuate, these fluctuations are inefficient. Randomness in θ implies randomness in consumption, only making the risk-averse representative household worse off. One way to resurrect a role for monetary policy in this model might be to add a shock to productivity that is not learned until after the representative household has chosen x. In this case, conducting open market operations to vary the quantity of liquidity in the asset (labor) market in response to the technology shock would be efficient for the central bank. The conjecture is that an optimal policy would involve "leaning against the wind" by injecting more liquidity when the technology shock is high, which would relax the household's second cash-in-advance constraint when the demand for labor is high. This rationale for monetary policy relies on the central bank being capable of acting faster than private agents to increase liquidity in the asset market. As well, this rationale relies on cash-in-advance producers, which is problematic, as discussed above.

Perhaps a more plausible approach to the nonneutrality of money and liquidity effects in this vein is taken in Williamson (2004a), in a model where cash-in-advance constraints are derived endogenously from first principles, and these cash-in-advance constraints apply to purchases of retail and wholesale goods. Credit is permitted so that the results do not depend on all purchases being made with outside money. A key result in Williamson (2004a) is that permitting private intermediaries to issue close substitutes for government-provided outside money alters the nature of cash-in-advance constraints, takes away the liquidity effect, and substantially changes optimal monetary policy rules.

3. LIMITED PARTICIPATION AND THE DISTRIBUTIONAL EFFECTS OF MONETARY POLICY

Much of the literature on limited participation and monetary policy has focused on liquidity effects in asset markets (e.g., Lucas [1990], Alvarez and Atkeson [1997], Alvarez, Atkeson and Kehoe [2002]) while neglecting the implications of limited participation for the distributional effects of monetary policy on output, consumption, and wealth. The heterogenous agent models studied by Grossman and Weiss (1983) and Rotemberg (1984) captured some of these distributional effects but not in a tractable way. Some economic agents receive the first-round impacts of monetary policy actions while others do not, making a difference for the distribution of wealth and for production and consumption across economic agents. Indeed, these distributional effects may be very important for how monetary policy works, if not *the* reason we should care about monetary policy.

In focusing on asset pricing implications, those working in the limited participation literature have also paid scant attention to normative issues. As I have shown in the previous two sections, my baseline model (which captures the key results in the literature) does not have much to say about how to conduct monetary policy. Focusing almost exclusively on optimal monetary policy, as I do here, will be helpful in showing how interesting policy conclusions arise when we are serious about modeling the distributional effects of monetary policy. The model here can also be used to explore the dynamic effects of monetary injections on output, prices, consumption, and interest rates, but that would turn this into a much longer article than the editor would allow.

In this section, I will modify the baseline model to incorporate distributional effects of monetary policy. To cleanly focus on these effects, I will leave out discussion of asset pricing implications. Most of the ideas in this section come from Williamson (forthcoming), but the model studied there is a monetary search model that builds on Lagos and Wright (forthcoming). The Lagos-Wright model is an approach to handling the distribution of wealth in search models with monetary exchange through the use of quasi-linear preferences rather than a representative household. I can capture the same ideas here as in Williamson (forthcoming) by extending my baseline cash-in-advance model. Though typical cash-in-advance models lack the microfoundations that make monetary search models such as Lagos and Wright (forthcoming) and Williamson (forthcoming) attractive, it is possible to generate cash-in-advance constraints while remaining true to monetary fundamentals, as, for example, in Williamson (2004a).

Suppose now that the representative household consists of two shoppers, two workers, and a continuum of financial transactors with mass 2, and that two locations, denoted location 1 and location 2, exist. One shopper and one worker live at each location, and a unit mass of financial transactors is always at each location. At the beginning of the period, a unit mass of financial transactors arrives at each location to deliver beginning-of-period money balances. In location 1, the shopper buys goods from other households on credit. That is, the shopper exchanges IOUs for goods, and the IOUs are redeemed by the household at the end of the period. At location 2, shoppers buy goods with money. The worker at each location sells goods in exchange for IOUs at location 1 and for money at location 2. At the end of the period, after IOUs clear, the financial transactors take possession of the household's money balances. Financial transactors are then randomly allocated (by nature) to each location. A financial transactor who is at a given location at the end of the period will be at the other location with probability π and in the same location with probability $1 - \pi$, where $0 < \pi < 1$. Given the random relocation of financial transactors, it is optimal for the household to allocate its money balances equally among financial transactors within a location.

The reason the financial transactors play the role they do in this version of the model is to have a convenient device for allowing money to diffuse through the economy. I could have accomplished a similar goal by having economic agents randomly allocated to the two locations to buy or sell goods. The key feature of the model I need to achieve my results is some friction associated with moving money and goods across locations. The exact form this friction takes is not so important, and for my purposes it is convenient that producers and consumers cannot move and that the household can move money across locations, though in a random fashion.

For convenience I have included only one asset—money—in this version of the model, so I cannot model central bank intervention as open market operations. Here, the government injects money into the economy through lump-sum transfers, which, for my purposes, is harmless in that the policy implications should not be qualitatively different from what I would get with a pure monetary policy experiment. The household receives the transfer at location 1 before financial transactors are randomly relocated. It is a key feature of the model that only some agents (those at location 1) receive the money transfer. Note also that the transfer is received by agents who have access to the more sophisticated transactions technology that involves within-period credit, capturing the fact that central bank intervention occurs in markets where financial transactions are relatively more complex than in other sectors of the economy.

In this version of the model, it will be interesting to explore optimal monetary policy in the context of aggregate shocks to the economy, so I will add an aggregate technology shock. Assume that one unit of labor produces ϕ_t units of the consumption good in period t, where ϕ_t follows a first order Markov process. I did not consider technology shocks in the previous versions of the model because the implications of doing so would be no different from those obtained in standard cash-in-advance models. Studying the behavior of the economy under technology shocks will yield important new results in this instance.

Since consumption goods cannot be moved between locations, consumption will in general differ between the two locations. It will be useful to suppose that the shoppers in the household do the consuming, with c_{it} denoting consumption by the shopper at location i. Similarly, n_{it} denotes labor supply by the worker at location i. Then, the household maximizes

$$E_0 \sum_{t=0}^{\infty} \beta^t \left[u(c_{1t}) + u(c_{2t}) - v(n_{1t}) - v(n_{2t}) \right]. \tag{36}$$

The household must abide by its budget constraint at location 1,

$$P_{1t}c_{1t} + M_{1,t+1} \leq P_{1t}\phi_t n_{1t} + (1-\pi)M_{1t} + \pi M_{2t} + \Upsilon_t, \tag{37}$$

where P_{it} denotes the price of goods in terms of money at location i, M_{it} is the quantity of money held by the household at location i at the end of period $t-1$, and Υ_t is the lump-sum money transfer from the government. As well, the household must satisfy the cash-in-advance constraint at location 2,

$$P_{2t}c_{2t} \leq \pi M_{1t} + (1-\pi)M_{2t}. \tag{38}$$

Finally, the household faces its budget constraint at location 2,

$$P_{2t}c_{2t} + M_{2,t+1} \leq P_{2t}\phi_t n_{2t} + \pi M_{1t} + (1-\pi)M_{2t}. \tag{39}$$

Let $\overline{M}_{it+1}$ denote the money supply in location i in period t after the government executes the transfer and before financial transactors are relocated. Then

$$\overline{M}_{1,t+1} = (1-\pi)\overline{M}_{1t} + \pi\overline{M}_{2t} + \Upsilon_{t+1}, \text{ and} \tag{40}$$

$$\overline{M}_{2,t+1} = \pi\overline{M}_{1t} + (1-\pi)\overline{M}_{2t}. \tag{41}$$

As in the previous sections, I can write the household's optimization problem as a dynamic program with analogous notation, except that the scaling variable I use is $\overline{M}_{2t}$, the quantity of money in location 2. Let z_{it} denote the gross growth rate in the money stock in location i, and let $V(m_1, m_2, \phi, z_2')$ denote the household's value function. Note here that it is sufficient to include only the money growth factor in location 2 in the state vector. Then, from (36) through (41), the household's dynamic programming problem is

$$\begin{aligned} V(m_1, m_2, \phi, z_2') &= \max_{c_1,c_2,n_1,n_2,m_1',m_2'} u(c_1) + u(c_2) - v(n_1) - v(n_2) \\ &\quad +\beta E_t V(m_1', m_2', \phi', z_2'')], \end{aligned} \tag{42}$$

subject to

$$p_1 c_1 + z_2' m_1' \leq p_1 \phi n_1 + (1-\pi) m_1 + \pi m_2 + \tau, \tag{43}$$

$$p_2 c_2 \leq \pi m_1 + (1-\pi) m_2, \text{ and} \tag{44}$$

$$p_2 c_2 + z_2' m_2' \leq p_2 \phi n_2 + \pi m_1 + (1-\pi) m_2. \tag{45}$$

Then, assuming that the cash-in-advance constraint (44) binds, and given the equilibrium conditions $\pi m_1 + (1-\pi) m_2 = 1$ (money demand equals money supply at location 2) and $c_i = \phi n_i$ for $i = 1, 2$ (the demand for consumption goods equals the supply at each location), the first order conditions from the optimization problem on the right-hand side of the Bellman equation (42) yield

$$u'(\phi n_1)\phi - v'(n_1) = 0, \tag{45a}$$

which solves for n_1. The first order conditions and the appropriate envelope conditions yield the two Euler equations

$$z_2' v'(n_2) n_2 = \beta E_t \left[\pi \psi' + (1-\pi) u'(\phi' n_2') \phi' n_2' \right] \text{ and} \tag{46}$$

$$z_2' \psi = \beta E_t \left[(1-\pi) \psi' + \pi u'(\phi' n_2') \phi' n_2' \right], \tag{47}$$

which then solve for (ψ, n_2) as a function of the state (ϕ, z_2'), where

$$\psi \equiv \frac{v'(n_1)}{\phi p_1}. \tag{47a}$$

No Aggregate Uncertainty

First, consider the case where there is no uncertainty about productivity; that is, ϕ_t is known at date 0 for all t. This setup is useful for studying how the central bank should behave in response to predictable events. While some of these predictable events—having to do with the day of the week, the month of the year, or the time until the end of the reserve-averaging period—are not necessarily appropriately modeled as related to aggregate productivity, the case of predictable productivity fluctuations will nevertheless be quite instructive.

When no aggregate uncertainty exists, then from (46) and (47) an equilibrium consists of sequences $\{\psi_t\}_{t=0}^{\infty}$, $\{n_{2t}\}_{t=0}^{\infty}$ that solve the difference equations

$$z_{2,t+1}v'(n_{2t})n_{2t} = \beta\left[\pi\psi_{t+1} + (1-\pi)u'(\phi_{t+1}n_{2,t+1})\phi_{t+1}n_{2,t+1}\right] \text{ and } \quad (48)$$

$$z_{2,t+1}\psi_t = \beta\left[(1-\pi)\psi_{t+1} + \pi u'(\phi_{t+1}n_{2,t+1})\phi_{t+1}n_{2,t+1}\right], \quad (49)$$

for $t = 0, 1, 2, ...$. The first important implication is that money is not neutral here because of a distribution effect. Suppose, for example, that productivity is constant, or $\phi_t = 1$ for all t, and that the stocks of money in locations 1 and 2 in period 0 are γM_0 and M_0, respectively, where $\gamma > 0$. After date 0, suppose that there are no transfers, so that the aggregate money stock is constant for all time. In typical monetary models, γ would have no effect on real aggregate variables; that is, money would be neutral. Here, $n_{1t} = n_1$ for all t, where n_1 is the solution to (45a) with $\phi = 1$. However, γ matters for the determination of $\{n_{2t}\}_{t=0}^{\infty}$, as from (40) and (41) γ will affect z_{2t} for each $t = 1, 2, ...$, which will, in turn, affect $\{\psi_t\}_{t=0}^{\infty}$ and $\{n_{2t}\}_{t=0}^{\infty}$ from (48) and (49). I will not go into detail here concerning the qualitative and quantitative nonneutralities of money in this model, as I want to focus in this section on optimal monetary policy, but these nonneutralities are potentially very interesting and worthy of study.

An optimal allocation is very easy to characterize in this model, as there is a single representative agent, and the optimization problem that an omniscient social planner would solve in this environment is a very simple static problem. That is, optimal n_{it}, for $i = 1, 2$, solves

$$\max_{n_{it}}\left[u(\phi_t n_{it}) - v(n_{it})\right].$$

Then, the first order condition for an optimum gives

$$\phi_t u'(\phi_t n_t^*) - v'(n_t^*) = 0. \quad (50)$$

Clearly, from (45a) and (50), employment is optimal in location 1, but employment will in general be suboptimal in location 2. From (48) through (50), the optimal allocation can be supported as a competitive equilibrium if $\{z_{2t}\}_{t=1}^{\infty} = \{z_{2t}^*\}_{t=1}^{\infty}$ where

$$z_{2,t+1}^* = \beta\frac{v'(n_{t+1}^*)n_{t+1}^*}{v'(n_t^*)n_t^*}. \quad (51)$$

The optimal allocation is then achieved in an equilibrium where

$$\psi_t = v'(n_t^*)n_t^* \text{ and}$$

$$p_{1t} = \frac{1}{n_t^*\phi_t}.$$

Now, if

$$-c\frac{u''(c)}{u'(c)} < 1, \tag{52}$$

so that the substitution effect dominates the income effect on labor supply for the household, then from (50), n_t^* is increasing in ϕ_t. Therefore, since $v'(n)n$ is increasing in n, the optimal money growth rate at location 2 in period t is increasing in ϕ_t and decreasing in ϕ_{t-1}. That is, the key monetary policy variable is the growth rate of the money stock in location 2, since location 2 is where transactions are conducted with outside money. At the optimum, monetary policy needs to correct for intertemporal price distortions due to (1) a suboptimal long-run rate of return on money and (2) the distortions introduced because output fluctuates in response to fluctuating productivity. To correct the first distortion, the money stock will tend to grow at the rate of time preference; note from (51) that if ϕ_t is constant for all t, then $z_{2,t+1}^* = \beta$ for all t. To correct the second distortion, since the price level will tend to be low when productivity and output are high (assuming (52)), money growth should be high when productivity is high.

The optimal money growth rule specified by (51) is typical of the optimal Friedman rules implied by representative-agent type monetary models in common use in macroeconomics. Friedman's (1969) prescription was to conduct monetary policy so that the nominal interest rate is zero in all states of the world. Though I have so far ignored the determination of nominal interest rates in this section, a standard approach to pricing a nominal bond would yield a zero nominal interest rate, given (51). Thus, so far nothing seems surprising about the implications for optimal monetary policy coming out of this model. However, the money growth rates specified by equation (51) are for the growth rates of the money stock in location 2 only. Note that the government controls these money growth rates only indirectly, through monetary intervention in location 1. It would be useful to see what (51) implies for the behavior of the money stock in location 1. Using (40), (41), and (51), the optimal money growth rates in location 1 are given by

$$z_{1t}^* = \left(\frac{z_{2,t+1}^* + \pi - 1}{z_{2t}^* + \pi - 1}\right) z_{2t}^*. \tag{53}$$

The optimal money growth rule given by (51) and (53) is more complicated than the simple Friedman rule in (51). This is because the indirect control of the money stock at location 2 through money injections and withdrawals at location 1 requires that the monetary authority take account of how the pattern of transactions diffuses money through the economy. In particular, note the role of π in determining the optimal money growth rate in location 1, where π governs the speed of diffusion of money through the economy. If $\pi = \frac{1}{2}$, then diffusion occurs in one period, while there is no diffusion if $\pi = 0$ or

$\pi = 1$. The speed of diffusion increases with π for $0 < \pi < \frac{1}{2}$ and decreases with π for $\frac{1}{2} < \pi < 1$.

Aggregate Uncertainty

It proves to be quite easy to generalize the optimal monetary rule given by (51) and (53) to the case where ϕ_t is an arbitrary first order Markov process. Here I want to consider how the monetary authority should react to unanticipated shocks to productivity that may be serially correlated. As in the previous subsection, a social optimum is $n_{it} = n_t^*$, for $i = 1, 2$, where n_t^* is the solution to (50). Then, from (46) and (47), an optimal money growth rule is given by

$$z_{2,t+1}^* = \beta \frac{E_t \left[v'(n_{t+1}^*) n_{t+1}^* \right]}{v'(n_t^*) n_t^*}, \tag{54}$$

and, as before, given the optimal money growth factor for location 2 from (54), the optimal money growth factor for location 1 is specified by (53).

Similar to the previous subsection, the optimal money growth rule specified by (54) and (53) has features similar to a standard Friedman rule in that the money stock at location 2 grows at the rate of time preference, modified by the corrections necessary for anticipated optimal growth in real output. Also, the optimal rate of growth in the money supply at location 1 follows a much more complicated rule for the same reasons as discussed in the previous subsection.

It may seem puzzling that the monetary authority can manipulate the money supply to achieve an optimal allocation, in spite of the fact that production and consumption occurs in two locations and the monetary authority can intervene directly only in one location. Critical to this result is that monetary exchange occurs only at location 2, that the important monetary variable is next period's money growth rate at location 2, and that the monetary authority can control that variable perfectly through current transfers at location 1. An interesting extension of this framework, which relates to my current research, is to allow for monetary exchange in both locations. In that case, the prices at which money trades for goods will in general differ across locations, and Friedman rules cease to be optimal monetary policy. This extension is much harder to study but could be potentially very fruitful for thinking about the role of monetary policy in actual economies.

4. CONCLUSION

While limited participation asset-pricing models such as the one studied by Lucas (1990) provide an explanation for the liquidity effect of monetary policy on nominal interest rates, these models do not provide a rationale for central banking or any guidance as to how a central bank should behave. Extensions of these models, such as in Fuerst (1992), that allow for nonneutralities of money lack plausibility, as they constrain firms to use cash in situations where their real-world counterparts use credit. In the latter part of this article I explored an extension of limited participation models that takes seriously the idea that monetary policy matters in the short run through its effects on the distribution of wealth across the population.

In ongoing research, I intend to explore further the qualitative and quantitative implications of a related class of limited participation models for monetary policy. These models represent a serious alternative to the sticky-price and sticky-wage Keynesian models that have been popular in recent policy analysis.

REFERENCES

Alvarez, Fernando, and Andrew Atkeson. 1997. "Money and Exchange Rates in the Grossman-Weiss-Rotemberg Model." *Journal of Monetary Economics* 40 (3): 619–40.

______________, and Patrick Kehoe. 2002. "Money, Interest Rates, and Exchange Rates with Endogenously Segmented Markets." *Journal of Political Economy* 110 (1): 73–112.

Christiano, Lawrence J., and Martin Eichenbaum. 1995. "Liquidity Effects, Monetary Policy, and the Business Cycle." *Journal of Money, Credit, and Banking* 27 (4): 1113–36.

Fuerst, Timothy S. 1992. "Liquidity, Loanable Funds, and Real Activity." *Journal of Monetary Economics* 29 (1): 3–24.

Grossman, Sanford, and Laurence Weiss. 1983. "A Transactions-Based Model of the Monetary Transmission Mechanism." *American Economic Review* 73 (5): 871–80.

Lagos, Ricardo, and Randall Wright. Forthcoming. "A Unified Framework for Monetary Theory and Policy Analysis." *Journal of Political Economy.*

Lucas, Robert. 1990. "Liquidity and Interest Rates." *Journal of Economic Theory* 50 (2): 237–64.

Rotemberg, Julio J. 1984. "A Monetary Equilibrium Model with Transactions Costs." *Journal of Political Economy* 92 (1): 40–58.

Shi, Shouyong. 2004. “Liquidity, Interest Rates, and Output.” University of Toronto, Department of Economics Working Paper 03-06.

Williamson, Stephen D. 2004a. “Limited Participation, Private Money, and Credit in a Spatial Model of Money.” *Economic Theory* 24 (November): 857–76.

______________. Forthcoming. “Search, Limited Participation, and Monetary Policy.” *International Economic Review.*

Bank Risk of Failure and the Too-Big-to-Fail Policy

Huberto M. Ennis and H. S. Malek

There seems to be a perception among participants in U.S. financial markets that if a large banking organization were to get in trouble, the government would, under most circumstances, intervene to prevent its failure (or limit the losses to uninsured creditors upon failure). This possibility of a government bailout is commonly referred to as the "too-big-to-fail" policy. The idea behind this belief is that, in general, policymakers will be inclined to bail out institutions which are considered to be of "systemic" importance; that is, institutions whose potential failure could threaten the stability of the entire financial system.

The expectation of contingent bailouts tends to create efficiency costs in the economy. In general, a bank tends to become larger and riskier if its uninsured creditors believe that they will benefit from too-big-to-fail (TBTF) coverage. In this article we provide a formal discussion to clarify the origin of these distortions and review empirical evidence on the relative importance of these distortions in the U.S. banking system.

The TBTF subject is a timely issue. Stern and Feldman (2004) argue that the problem of TBTF is actually getting worse. They identify the increasing concentration and complexity in banking as the main reason for this deterioration. Although their opinion is certainly not shared by everyone, the mere possibility of such a costly distortion is enough to justify further study of this issue.

The too-big-to-fail terminology sometimes can be misleading. While the systemic importance of an organization tends to be closely related to its size, this is not always the case. For example, a handful of U.S. banks are not

We would like to thank Allen Berger, Tom Humphrey, Leo Martinez, Ned Prescott, Jennifer Sparger, and John Weinberg for comments on a previous draft and Kevin Brown, Ross Waldrop, and John Walter for their help with the data. E-mails: Huberto.Ennis@rich.frb.org; Sam.Malek@rich.frb.org. The views expressed here do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

particularly large but are still often perceived as too big to fail because they perform an essential activity in the smooth functioning of financial markets and the payment system. Furthermore, the TBTF problem is not exclusive to banks. Other financial intermediaries like large clearinghouses and significant players in the mortgage securities market are often perceived as too big to fail. In this article, however, we will restrict our focus to traditional banking activities and, for simplicity, will consider size as the main variable associated with the likelihood of being bailed out.

U.S. banks face a complex regulatory environment that guides and modifies their behavior. The perception of a TBTF policy is just one of several features that characterizes this environment. Two other important features tend to interact with TBTF: deposit insurance and the failure-resolution policy.[1]

The Federal Deposit Insurance Corporation (FDIC) is an independent government agency that provides deposit insurance to U.S. banking institutions. The current insurance system protects a depositor's insured funds up to $100,000, including principal and interest. The FDIC administers two insurance funds: the Bank Insurance Fund (BIF), which is dedicated to commercial banks, and the Savings Associations Insurance Fund (SAIF) for the savings and loans banks. Member-banks contribute periodic payments to a common pool, which is then used to finance the insurance liabilities in case of a bank failure. Prior to 1993, all banks paid to the FDIC the same contribution per dollar of deposits. However, since 1993, the contributions are partially based on risk. Under this new system, institutions are grouped into nine risk categories according to their level of capitalization and the rating obtained during supervisory examinations. Banks belonging to the higher risk categories are required to pay higher premiums. The range of premiums is updated semiannually by the FDIC according to the funding needs of the insurance funds. Presently, the premiums range from 0 to 0.27 percent of deposits. Since 92 percent of banks satisfy the requirements for a 0 percent assessment, they do not contribute to the fund. The target size of the fund is 1.25 percent of total insured deposits in the system, and, in case of unexpected financial pressure, the current regulation allows for the fund to draw on a $30 billion line of credit from the U.S. Treasury (to be repaid with future premiums by member banks).

As part of a response to a pronounced crisis in commercial banking resulting in a BIF deficit of $7 billion, Congress passed the FDIC Improvement Act (FDICIA) in December 1991.[2] The Act introduced risk-based premiums and new regulations for bank-failure resolution. The new rules specify a course of action for regulators to enforce adjustments in undercapitalized banks and, in this way, mitigate the potential losses to the fund associated with bank failures. Before FDICIA, the power to close a failing insured bank rested

[1] See Hetzel (1991) for a discussion of TBTF and the timely closing of insolvent banks.

[2] For a comprehensive survey of FDICIA, see Benston and Kaufman (1997).

with the chartering authority (either the Comptroller of the Currency or state governments). Nowadays, an institution whose capital ratio falls below 2 percent faces closure by the FDIC if the shortfall is not corrected within 90 days (see Walter [2004] for details). While the regulatory reforms introduced in FDICIA limit the protection of uninsured creditors, Section 141 still considers the possibility of a TBTF bailout. This "systemic risk" exception attempts to increase scrutiny over bank bailouts by requiring that both the Federal Reserve and the Treasury sign off on a rescue.[3]

Evidently, the complex deposit insurance system—in combination with the potential for TBTF coverage—creates an intricate set of incentives that influences the decisions of U.S. banks. In the model we provide to analyze the banks' decision process, banks are competitive and must offer the best possible contract to attract potential creditors. We show that when the deposit insurance system involves premium payments that do not fully reflect risk, banks tend to become riskier to exploit the potential net transfer to their creditors under the contingency of failure. We also study partial coverage and the interaction between deposit insurance and a TBTF policy. In particular, we show that the TBTF policy creates not only a risk distortion but also a size distortion, and that one distortion tends to increase the value of the other (and vice versa), creating a perverse amplification effect.

We model risk in a simple yet useful way. We consider only the risk of failure in the decision of banks. This simplification is appropriate for the study of TBTF, which is linked only to the events in the distribution of outcomes that result in failure. Of course, in general, the risk of failure is a consequence of a set of risky decisions made by banks. These decisions also imply a complex distribution of returns when the bank does not fail. We abstract from this aspect of the risk involved in banking and assume that if the bank does not fail, it has a fixed return R.

Studying the cost and benefits of TBTF bailouts is difficult. Failures of large banks are low-probability events. As a consequence, we do not have sufficient data to fully identify the pattern of behavior (of bankers, policymakers, and creditors) linked to bailouts. Also, the indirect (moral hazard) effect of TBTF on the investment portfolio of banks is difficult to discern. At the same time, the decision to bail out a particular bank depends on a large number of circumstances, and reaching general conclusions based on specific events is not good practice. For example, observing that a relatively important failing bank is not bailed out may help elucidate the position of policymakers

[3] Regulators often argue that even if a troubled financial institution is not closed, this does not mean that all its major claimants are protected from losses. In general, the regulators of a troubled institution might have its management removed and its existing equity extinguished. Also, sometimes significant (partial) losses might be imposed on uninsured creditors and counterparties (Greenspan 2000). Clearly, all these instruments will contribute to limit the distortions created by the perception of a TBTF policy.

with respect to the TBTF policy. However, just one situation is probably not enough evidence to conclude that TBTF is not a problem. A different bank, in different situations, may actually be bailed out. In other words, it may be useful to think about the bailout event as probabilistic, which is the approach that we take in this article. In the next section, we present a model where the probability of a TBTF bailout is strictly between zero and one (for a relevant set of bank sizes), and (on this range) such a probability is increasing in the size of the bank.

In the second section of this article, we revisit some empirical evidence first presented by Boyd and Gertler (1994), who studied the relationship between bank performance and asset size in the United States and concluded that the evidence indicates the emergence of a TBTF problem in the late 1980s. We extend that analysis to the period 1991–2003, revealing that the patterns justifying Boyd and Gertler's concerns are no longer in the data. We provide some interpretations for this change.

It is important to point out that we are not discussing why a TBTF policy may be in place. Rather, we assume that there is a TBTF policy and then identify its potential effects on the size and risk decisions of banks. This assumption simplifies the exposition and allows us to focus exclusively on the distortions introduced by TBTF. But the simplification does not come without cost. In particular, we do not discuss two important issues related to the existence of TBTF bailouts: the potential benefits of avoiding spillovers and bank runs and the time inconsistency problem faced by policymakers. We refer the interested reader to the excellent discussion in Chapter 2 of Stern and Feldman (2004). However, we would like to stress here that we consider the study of those issues essential for a full understanding of the TBTF problem.

The remainder of the article is organized as follows. In Section 1 we present a simple model of the size and investment decision of competitive banks and study this decision under different explicit and implicit deposit insurance schemes. The model allows us to identify the distortions that the different possible schemes create on the level of risk taken by banks and the size of their operations. In Section 2, we review empirical evidence aimed at determining if the U.S. banking system functions under the perceptions of an implicit TBTF government insurance scheme. The last section provides concluding remarks.

1. A SIMPLE MODEL

Consider an economy with a large number of banks and a large number of agents that play the role of potential depositors. Each agent has 300 units of funds available, and they can either deposit some (or all) of their funds at a bank or invest them in a safe asset which provides a gross rate of return, given by r. The banks make risky investments and may fail with a certain probability, π. In the case that a bank does not fail, depositors get R units per unit deposited at that bank. We assume that R can take values in the interval $[0, \overline{R}]$, where $\overline{R}$ is an upper bound of the set of possible gross returns on bank deposits. Furthermore, we assume that banks can charge a fee, F, to each depositor.

Assume that the probability of bank failure, π, is increasing in R. This assumption captures the idea that taking higher risks is necessary to obtain higher returns. For simplicity, we assume that $\pi(R)$ is linear in R with slope, a. When the bank fails, we assume that no resources are left at the bank to pay depositors. In other words, without government intervention, depositors will get zero from the bank in case of failure. For reasons that will become clear shortly, we assume that r and $\overline{R}$ satisfy the following conditions:

$$r = \frac{1}{4a}, \text{ and } \frac{3}{4a} < \overline{R} < \frac{1}{a}.$$

Also, for simplicity we will assume that depositors can deposit an amount, x, of funds in the bank, where x can take one of three possible values: 50, 100, or 300. Furthermore, all depositors want to have at least 50 units deposited at the bank. We do not model explicitly the reasons for this minimum deposit, but the idea is that all agents wish to have at least some bank balances for settlement of "essential" payments.[4]

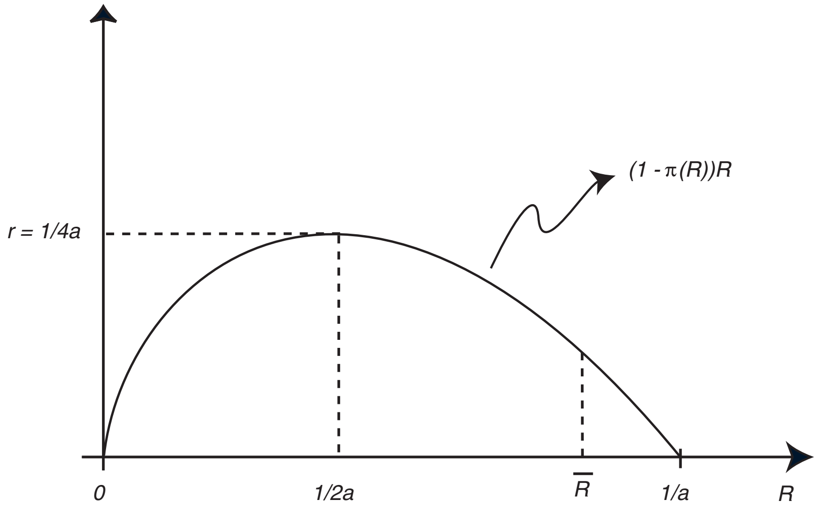
Finally, banks can choose their size. Let ξ be the proportion of the total population of agents making a deposit in a particular bank. To make the choice of ξ interesting, we assume the cost c per depositor of running a bank is convex in ξ with a minimum at ξ^o. The idea behind this assumption is that an optimal size of operation for banks exists and is associated with the size ξ^o. Running a bank that is too small (i.e., smaller than ξ^o) increases the operational cost per depositor; and running a bank that is too large (i.e., larger than ξ^o) also increases the cost.

We assume that banks compete to attract depositors. In equilibrium, banks earn zero profits and choose R and ξ so as to make the expected payoff to a depositor as high as possible. If a bank were not to follow such a strategy, some other bank would arrange its choices of R and ξ in order to attract all the depositors from the first bank. This equilibrium concept is standard in the banking literature. All agents and banks are identical, and in equilibrium they behave symmetrically. As a consequence, the equilibrium value of ξ is a good proxy for the size of the representative bank.

We now study different banking arrangements and their effects on the risk of failure and the size chosen by the banks.

[4] The discreteness in the size of deposits is assumed only for the sake of simplicity. It allows us to capture the main reasons driving agents' decisions without complicating the calculations.

Figure 1 Optimal Return-Risk Combination



Notes: In a laissez-faire system, banks set the return $R^L = 1/2a$, which maximizes the expected return per unit deposited (net of fees) given by $(1 - \pi(R))R$.

Laissez-faire System

Consider first the case of a laissez-faire banking system—that is, one without any government intervention. The laissez-faire equilibrium provides an important benchmark for our evaluation of alternative explicit and implicit deposit insurance systems in the following subsections. Under laissez faire, the expected payoff to a depositor is given by

$$(1 - \pi(R))\, xR + \pi(R)0 - F,$$

where the equilibrium fee, F, will cover the operational costs per depositor, $c(\xi)$. Let us call R^L and ξ^L the laissez-faire equilibrium values of R and ξ. These values maximize the payoff to depositors and, hence, must satisfy the following necessary and sufficient conditions:

$$\frac{d\pi}{dR} xR^L - \left(1 - \pi(R^L)\right) x = 0,$$

and

$$\frac{dc(\xi^L)}{d\xi} = 0,$$

which imply that R^L equals $1/2a$ and ξ^L equals ξ^o. Note that R^L is the value of R that maximizes the payoff, $(1 - \pi(R))xR$ (see Figure 1).

To complete the analysis, we need to determine if the depositors would find it beneficial to deposit in these banks any amount in excess of 50 units. If an agent deposits the minimum 50 units of its funds in a bank and the remaining 250 in the safe asset, then its expected payoff will be given by

$$\left(1 - \pi(R^L)\right) 50R^L - c(\xi^L) + 250r.$$

We need to compare this alternative with that of depositing any other feasible amount, x, greater than 50 (in particular, $x = 100$ or 300). The net benefit of increasing the amount deposited at a bank to $x > 50$ is given by

$$\left(1 - \pi(R^L)\right) (x - 50)R^L - (x - 50)r.$$

Recall that we assumed that $r = 1/4a$. Then, since $(1 - \pi(R^L))R^L = 1/4a$, we obtain that the net benefit is zero, and for any amount in excess of 50, depositors would be indifferent between making an investment or a deposit.

It is important to note that the model presented here has no inherent interaction between size and risk, even though in reality there may be reasons to believe that a bank's size and risk of failure can be associated in some fundamental way. This simplification is useful because it allows us to concentrate on the interactions between size and risk that may originate in specific banking policies.

Deposit Insurance

We will consider four different deposit insurance systems. The systems differ from one another in the structure of premiums and the coverage that they provide.

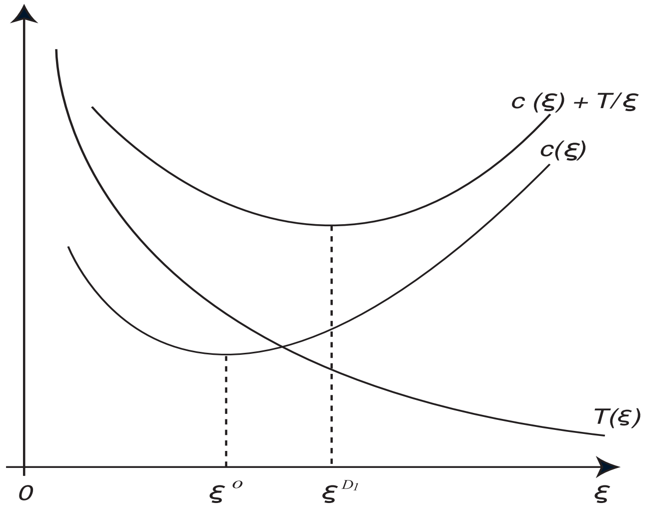
We start with a deposit insurance system that provides full coverage of losses and in which banks pay to the insurance fund a lump-sum fixed premium, T, independent of bank size. While this kind of fixed premium seems unrealistic, such an extreme assumption is useful to illustrate how misalignments in the premium structure can create size distortions. In this simple model, designing the right premium structure to avoid this kind of size distortion is straightforward, and we describe such a structure below.

Under this system, banks choose the values of R and ξ that solve the following problem:

$$\max_{R,\xi}\, (1 - \pi(R))\, xR + \pi(R)xR - F, \tag{1}$$

where $F = c(\xi) + T/\xi$. Let us denote the solution to this problem with (R^{D1}, ξ^{D1}). It is then clear that under full coverage the banks will choose $R^{D1} = \overline{R}$, the maximum value of the possible (risky) returns. Recall that the probability of failure of a bank is increasing in R and, hence, by setting R^{D1} equal to $\overline{R}$, banks will be indirectly maximizing the probability of failure. Banks follow this strategy because the insurance premium that a bank pays

Figure 2 Optimal Bank Size



Notes: The size of a bank that minimizes the per-depositor cost of operation $c(\xi)$ is given by ξ^o. When the bank is paying a fixed lump-sum premium T to the deposit insurance fund, it will increase its size to ξ^{D1}.

does not depend on the risk taken by the bank, and, furthermore, full insurance coverage implies that neither the bank nor its depositors face any downside from selecting higher levels of risk.

With respect to the equilibrium size of banks, we have that the value of ξ that maximizes the objective in problem (1) solves the following (first order) condition:

$$\frac{dc(\xi)}{d\xi} = \frac{T}{\xi^2} > 0.$$

It is then straightforward to see that $\xi^{D1} > \xi^o$. Recall that ξ^o was the size of the bank that minimizes the cost of operation $c(\xi)$. Here, however, by becoming large, the bank reduces the per capita cost of deposit insurance for depositors. Hence, the optimal size of the bank is larger than the one that minimizes operational costs. In other words, the lump-sum premium distorts the optimal-size decision by banks (see Figure 2).

To avoid the size distortion, the deposit insurance fund could make the premium, T, dependent on the size of the bank. This structure of premiums makes sense to the extent that, for a given level of risk, larger banks will impose higher costs to the insurance fund. Suppose, for example, that $T = b\xi$. Then, it is straightforward to show that the bank will choose to be of the optimal size ξ^o.

While this kind of premium scheme will solve the size distortion, there still remains the risk distortion. In fact, under this structure of premiums, banks would still choose to maximize the probability of failure. Of course, the fund could implement alternative regulations to limit the amount of risk taken by banks. For example, it could restrict the types of investments allowed to banks so that the bank would not be able to choose a level of R as high as $\overline{R}$. However, this model is too simple to study these more sophisticated regulations.

One other possibility would be to make the premium contingent not only on size, but also on risk. In fact, by choosing T to equal $\pi(R)xR\xi$, the insurance fund would give banks the necessary incentives to choose $R = R^L$, the same rate that banks would choose under laissez faire. In general, though, precisely assessing the risk taken by banks is difficult, and we can expect that the observed premium payments will not fully correct the risk distortion introduced by deposit insurance (Prescott 2002). For simplicity, in what follows we will assume the extreme case in which the premium only corrects the size distortion and is given by $T = b\xi$.

The last feature of deposit insurance that we wish to study is partial coverage. To be precise, suppose that in the case when a bank fails, the deposit insurance fund covers only up to 100 units of funds per depositor. Then, banks will choose the risk and size that solve the following problem:

$$\max_{R,\xi} (1 - \pi(R))\, xR + \pi(R) \min\{x, 100\}R - c(\xi) - b.$$

Let us call the solution (R^{D2}, ξ^{D2}). Since the total premium, T, is increasing with size, there will not be a size distortion in the decision of banks and therefore, ξ^{D2} equals ξ^o. With respect to the level of risk-return, R, the choice of banks will depend on whether the typical depositor has more or less than 100 units deposited at a bank.

For $x \leq 100$, the insurance provided is effectively full insurance. Then, as we saw before in the full-coverage case, depositors would find it most beneficial if banks maximize the risk-return combination.

Only if depositors have $x > 100$ does the partial coverage provide incentives to reduce risk at banks. In the banking literature, these depositors have been named "uninsured depositors." This terminology is not completely precise to the extent that all depositors receive insurance for the funds below the 100 limit. However, the terminology does convey the idea that these depositors are the ones susceptible to the risk of failure of their bank.

The coverage limit helps reduce the risk distortion but in general will not be enough to fully correct it. To see this, suppose that the typical depositor deposits 300 units of funds at the bank. Then, the bank will choose a level of

R that solves the following first order condition:

$$\frac{d\pi}{dR}R^A - \left(1 - \pi(R^A)\right) = \frac{1}{3}\left[\frac{d\pi}{dR}R^A + \pi(R^A)\right] > 0. \tag{2}$$

Recall that $R^L = 1/2a$ is the value of R that makes the left-hand side of equation (2) equal to zero (see Figure 1). Hence, since the right-hand side of this equation is positive, R^A must be greater than R^L, and the risk distortion is still present. For most cases, R^A will be smaller than $\overline{R}$, and we can say that, in the presence of uninsured depositors, the insurance limit can partially resolve the risk distortion introduced by deposit insurance.[5]

From the previous discussion we can then conclude that R^{D2} is either equal to R^A (if $x > 100$) or to $\overline{R}$ (if $x \leq 100$) and, hence, greater than R^L in either case. To determine the actual value that R^{D2} will take in equilibrium, we need to establish whether the typical depositor would be willing to deposit more than 100 units in a bank. The payoff from depositing more than 100 units is given by

$$\left(1 - \pi(R^{D2})\right)300R^{D2} + \pi(R^{D2})100R^{D2} - c\left(\xi^o\right) - b.$$

Alternatively, suppose that the agent deposits only 100 units at a bank and invests the rest in the safe investment with return r. In this case, the payoff is given by

$$100R^{D2} - c\left(\xi^o\right) - b + 200r.$$

Since $(1-\pi(R^{D2}))R^{D2} < r$ (see Figure 1), it is easy to see that depositing 100 units at a bank and the rest in the safe investment is the best strategy. Another alternative for the agent is to hold three deposit accounts of 100, each one at a different bank. This alternative will dominate both the 300-unit deposit and the alternative involving the safe asset described above. In fact, if a depositor can open any number of these accounts, then the 100-unit limit would never be relevant. It should be said, though, that opening accounts in several different banks involves transaction costs that are not being explicitly modeled here. One possibility for reducing these transaction costs is for the depositor to delegate this activity to a broker. However, in the U.S. system, brokered deposits are subject to regulations enforced by the supervisory agencies. For the sake of simplicity, in what follows we will assume that depositors can only have one bank account in the system.

Summarizing, the typical depositor in this banking system will have only deposits for 100 units or less, and banks will choose $R^{D2} = \overline{R}$—that is, the rate of return that corresponds to the highest feasible risk of failure. In other words, even though partial coverage has the potential for limiting risk-taking

[5] If x is greater than 100 (but less than 300), R^A may still equal $\overline{R}$. Here, then, the discreteness of the size of deposits simplifies calculations.

behavior by banks, it also creates incentives for depositors to stay below the limit, thereby undermining the disciplining mechanism.

Too Big to Fail

Suppose now that with probability, p, the bank is bailed out upon failure. To show that the bailout is spurred by the fear that a large organization's failure will disrupt the entire financial sector, we assume that p is increasing in the bank's size, ξ. This is a simple way to capture the too-big-to-fail policy. We still consider the case where a deposit insurance system with partial coverage is in place. Hence, the too-big-to-fail policy has consequences for the payoff of only those depositors with deposits above the limit. The payoff to depositors in the event of a bank failure is given by the function:

$$\Phi(R, \xi) \equiv \min\{x, 100\}R + p\,(\xi)\max\{0, x - 100\}R.$$

Competitive banks choose the values of R and ξ that solve the following problem:

$$\max_{R,\xi}\,(1 - \pi(R))\,xR + \pi(R)\Phi(R, \xi) - c\,(\xi) - b, \tag{3}$$

where the objective function is the expected payoff to the representative depositor. Let us call the solution to this problem (R^T, ξ^T). It is useful to start with the extreme case of banks that are so large that the probability of a bailout is unity (i.e., $p(\xi^T) = 1$). Then, problem (3) reduces to the full-coverage deposit insurance system we studied at the beginning of the previous subsection, and banks in equilibrium chose $R^T = \overline{R}$, which implies that the risk of failure would be maximized.

In the general case when the probability of bailout, p, is between zero and one, the solution to problem (3) suggests some interesting insights about the distortions introduced by the too-big-to-fail policy. This policy is relevant only for those agents that have uninsured deposits. Suppose then, that the typical depositor of the bank has $x > 100$. The partial derivatives of the payoff function, Φ, are given by:

$$\Phi_R(R, \xi) \equiv \frac{\partial\Phi(R, \xi)}{\partial R} = 100 + p\,(\xi)\,(x - 100)$$

and

$$\Phi_\xi(R, \xi) \equiv \frac{\partial\Phi(R, \xi)}{\partial\xi} = \frac{dp(\xi)}{d\xi}(x - 100)R;$$

and the solution (R^T, ξ^T) to the bank problem must satisfy the following first order conditions:

$$\frac{d\pi}{dR}xR^T - \left(1 - \pi(R^T)\right)x = \left[\frac{d\pi}{dR}\Phi(R^T, \xi^T) + \pi(R^T)\Phi_R(R^T, \xi^T)\right], \tag{4}$$

and

$$\frac{dc(\xi^T)}{d\xi} = \pi(R)\Phi_\xi(R, \xi). \tag{5}$$

Since $\Phi_R(R, \xi)$ and $\Phi_\xi(R, \xi)$ are both positive, $R^T > R^L$ and $\xi^T > \xi^o$. In other words, the too-big-to-fail policy induces banks to become larger and riskier than in a laissez-faire system. Furthermore, by comparing expression (4) with expression (2) (in the previous subsection) we see that, in general, R^T will be greater than R^A, which was the return chosen by a bank with uninsured depositors under no contingent-bailout policy.

One remaining question is whether depositors would want to deposit funds in excess of 100 in a banking system like the one we study in this subsection. The (net of fees) payoff to an agent depositing 300 units of funds at the bank is given by

$$\left(1 - \pi(R^T)\right) 300R^T + \pi(R^T)\,(1 + 2p\,(\xi))\,100R^T.$$

Comparing this payoff with the payoff from depositing only 100 units of funds (and the rest at the safe interest rate, r) we see that the difference is given by

$$\left[\left(1 - \pi(R^T)\right) R^T - r\right] 200 + \pi(R^T)p\,(\xi)\,200R^T. \tag{6}$$

Since R^T will generally be greater than R^L, we know that the first term in expression (6) is negative. However, the second term is positive, and for a large enough bailout-probability, p, it would compensate for the loss in the first term. It is then possible in this banking system for agents to find it beneficial to deposit all 300 units of funds at the bank.

Another interesting observation that results from expressions (4) and (5) is the interaction that exists between size and risk under the too-big-to-fail policy. Note that the right-hand side of expression (4) is increasing in p (which, in turn, is increasing in ξ). Then, the larger the bank, the larger the value of R the bank will wish to implement. Similarly, the right-hand side of expression (5) is increasing in R, and, hence, the higher the risk taken by a bank, the higher the incentives to increase its size. The reason for this complementarity between size and risk is that riskier banks are more likely to benefit from the possibility of bailouts (they are more likely to fail). Therefore, those banks are the ones that would like to increase the bailout probability, p, an objective that can be pursued by increasing the size of the banking organization.

This interaction captures the idea of a "virtuous circle" induced by an autonomous reduction on the probability of bailout (Stern and Feldman 2004, 21). Suppose that the appointment of a "conservative" regulator reduces the value of p for all values of ξ. This reduction in p will reduce the value R^T chosen by banks according to expression (4), which, in turn, will reduce the equilibrium size, ξ^T. A smaller ξ^T further lowers the risk taken by banks, reducing the failure probability and creating a virtuous circle that significantly reduces the likelihood of failure and bailout events.

As we have seen, the existence of a TBTF policy has two effects: it creates a size distortion in the banking industry, and it tends to accentuate the risk distortion that was already present under deposit insurance (i.e., R^T is greater than R^A). A commonly proposed policy to limit the effects of perceived implicit government guarantees is to limit the size of banks so that the probability p is equal to zero. Suppose, for example, that there is a bank size, ξ_p, such that $p(\xi) = 0$ for all $\xi \leq \xi_p$. Then, by limiting banks to be no larger than ξ_p, the government can eliminate the risk distortion originated in the TBTF perception. In general, however, limiting the size of banks will increase operational cost unless $\xi^o \leq \xi_p$. When the value of ξ is restrained by regulation to be below ξ_p, the value of R that banks choose solves a problem equivalent to the last problem studied in the previous subsection. It is somewhat ironic then that, in our model, limiting the size of banks to be smaller than ξ_p implies that banks will choose $R^{D2} = \overline{R}$, which could increase the riskiness in banking.

Another possible policy to limit the size of these distortions is to implement a system of "coinsurance" (Feldman and Stern 2004). The idea is that whenever a bank fails and gets bailed out, uninsured depositors will obtain only a proportion $\theta < 1$ of their deposits in excess of the insurance limit. The payoff in the event of a bank failure is now given by the function,

$$\Phi(R, \xi, \theta) = \min\{x, 100\}R + p(\xi)\,\theta \max\{0, x - 100\}R.$$

The bank problem is the same as in expression (3) but where $\Phi(R, \xi, \theta)$ replaces $\Phi(R, \xi)$. The solution to this problem will be a function of the parameter θ. Let us call such a solution (R^C, ξ^C). It is easy to see that for $\theta = 1$ we have $(R^C, \xi^C) = (R^T, \xi^T)$. However, for θ lower than unity, R^C is lower than R^T, and ξ^C is lower than ξ^T.[6] In other words, the coinsurance system reduces the incentives for banks to become bigger and riskier under a TBTF policy.

The deposit insurance premium, T, could be designed to reduce the size distortion induced by the TBTF policy. In particular, if the premium per unit deposited, b, is made increasing in the size of the bank, banks will have less incentive to become large, which, in turn, would limit the influence of the TBTF perception. The idea behind this strategy is important and can be restated in more general terms: whenever the TBTF problem is present, designing the structure of the deposit insurance premium to be neutral with respect to size (that is, in our model, $T = b\xi$) may not be optimal.

Finally, another way to control the risk-taking behavior of banks in the presence of a TBTF distortion is to directly limit the bank's activities via supervisory exams. In our simple model, this strategy amounts to reducing

[6] The solution R^C is lower than R^T as long as the coinsurance system does not make the optimal size of deposits equal 100 units or less.

the acceptable values of R that the bank may choose, or in other words, to lower the upper bound on returns, $\overline{R}$, a parameter in the model.[7]

2. THE ELUSIVE EVIDENCE

Boyd and Gertler (1994) look back at the banking troubles of the 1980s and find that "large banks were mainly responsible for the unusually poor performance of the overall industry" (p. 2). They attribute this feature of the data to the combination of two main factors: deregulation and too-big-to-fail. In particular, they argue that after the collapse of Continental Illinois Bank in 1984, it became clear that large banks were subject to a TBTF policy.[8] Using a panel of U.S. bank data for the period 1984–1991 they conclude that a robust negative correlation exists between size and performance and suggest that this correlation may be indicative of an increased perception of a TBTF subsidy.

The idea behind this strategy is that banks that are riskier ex ante, are also more likely to perform poorly ex post. Moreover, riskier banks, as a consequence of having more spread distribution of returns, tend to have a higher probability of failure.[9] Combining these two hypotheses implies that poorly performing banks have a higher probability of failure. Then Boyd and Gertler (1994, 15) postulate that "by examining ex post returns we can get some feel for the outer tails of the distributions." As we saw in the previous section, under the influence of a TBTF policy, banks will tend to increase the probability of failure. It is, of course, not obvious that increasing the probability of failure is always associated with an increase in the overall risk of the bank. Similarly, riskier banks do not always perform poorly, on average, relative to less risky banks. However, data limitations suggest that, in principle, the proposed link between risk, poor performance, and likelihood of failure may be a useful working strategy.

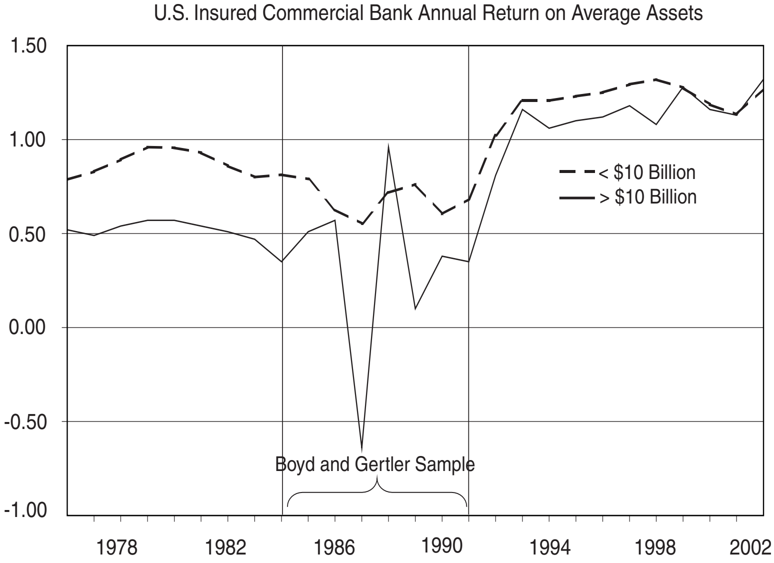
Boyd and Gertler use the decreasing trend in U.S. bank profitability during the 1980s as a starting point for their study. Specifically, they stress the fact that profitability was significantly below its 1970s average by the late 1980s. Our Figure 3 illustrates this fact. We plot the annual net income as a percentage of total assets for U.S. insured commercial banks. We divide banks in two groups, those with more than \$10 billion in total assets (large banks) and those with less than that amount. The decline in profitability during the 1980s is

[7] Our model does not allow us to study another form of controlling the risk-taking behavior by banks: capital requirements. See Prescott (2001) for a good formal introduction to the subject.

[8] In September 1984, the Comptroller of the Currency testified to the U.S. Congress that 11 bank holding companies were too big to fail (see O'Hara and Shaw 1990).

[9] In the previous section we did not allow for general distributions of returns which are an integral part of the interpretation for the evidence in this section. The link between the distribution of returns and the probability of failure is a technical issue that is not essential for understanding the incentives distortion introduced by TBTF, which was the main subject of the previous section.

Figure 3 Bank Performance and Size



Source: FDIC Quarterly Banking Profile.

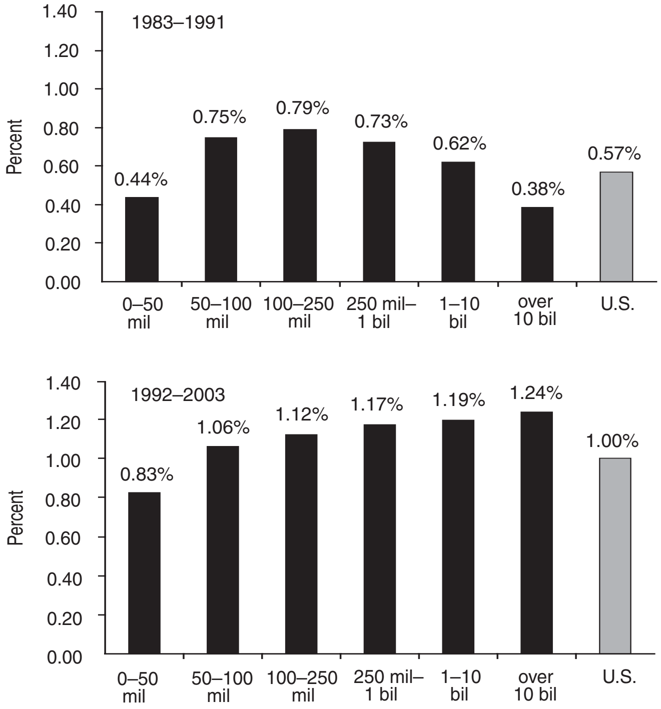
common for the two groups.[10] However, it is clear from the figure that large banks experienced an especially turbulent time during the second half of the 1980s. What is even more interesting is that after 1991, bank profitability recovered across the board to levels above those in the 1970s, staying fairly stable since then.

In summary, Figure 3 puts in perspective the sample period used by Boyd and Gertler and may cast some doubt on the robustness of their results. For this reason, we extend Boyd and Gertler's empirical analysis to include the data from 1992 to 2003.

Figure 4 presents the average return on assets for banks of different sizes. One of the main motivations for Boyd and Gertler's conclusions is the hump-

[10] Keeley (1990) argues that banks became riskier during the 1980s as a consequence of a generalized decrease in franchise value across the industry. Franchise value can help control risk-taking behavior by banks because bank owners fear losing this value upon failure. The evolution of banks' franchise value is an important determinant of their behavior, but, unfortunately, we will not have much to say about it in this article. See Demsetz, Saidenberg, and Strahan (1996) for further discussion of this issue.

Figure 4 Return on Assets and Size



Notes: We use data for all insured commercial U.S. banks (except credit card banks). To construct return on assets we divide annual net income by total assets. We consider each annual observation for each bank as the basic data entry in the calculation of averages across sizes (i.e., we do not take time averages for each bank). The total number of observations is around 120,000 for 1983–1991 and 110,000 for 1992–2003.

Sources: Report of Condition and Income Data (Call Report); Federal Reserve Bank of Chicago Web page.

shaped pattern of the first panel of Figure 4. Large banks performed relatively poorly during that period, presumably because of the improper pricing of risk induced by the TBTF distortion.[11] However, the second panel shows that in the period after 1991, the return on assets experienced by banks was, in fact, a monotone-increasing function of size. There are two competing explanations

[11] Banks with less that $50 million in assets also performed worse than the middle-sized banks. This pattern may be the consequence of the inability of small banks to exploit economies of scale.

for this change in pattern. Perhaps the hump-shaped pattern observed in the 1983–1991 period was the result of a special event at the end of the 1980s that hit hardest the performance of large banks. In this case, the pattern may not be related to a TBTF perception, and when no special event took place in the 1992–2003 period, the pattern disappeared. The other interpretation for the change in the pattern is that after 1991, changes on banking regulation and other policies induced a decrease in the likelihood of TBTF bailouts.

Before discussing the relevance of these two alternative explanations, we follow Boyd and Gertler's methodology and check whether the change in pattern just discussed is robust to controlling for regional effects. The idea behind this exercise is that the performance of banks may be driven by regional economic shocks. For example, if most of the large banks in the country are in a region that experienced an especially unfavorable shock during the period under study, then it is possible to find that, on average, mid-sized banks outperformed large banks just as a consequence of "location" effects.

While this type of "robustness" check may have been important for the 1983–1991 period, there are a priori reasons to believe that the adjustment is bound to be insignificant for the sample period of 1992–2003. First, several large banks today have nationwide operations and, hence, are less exposed to business fluctuations in specific regions. Second, looking at bank performance across regions during the 1992–2003 period does not reveal any clear regional disparities. The situation was not the same in the sample period studied by Boyd and Gertler, when the west-central region of the South and the west-central region of the Midwest experienced severe regional banking shocks.

Let us denote by D_j^r, a dummy variable indicating that a bank is headquartered in region j; by D_k^s, a dummy variable indicating that a bank belongs to size class k; and by x_{ijk}, a time-average value of bank return on assets. We run the following regression to obtain estimates of size effects on performance, controlling for a region:

$$x_{ijk} = a_j D_j^r + b_k D_k^s + \varepsilon_{ijk}.$$

This is equation (1) in Boyd and Gertler (1994). We construct two sets of time-average return on assets, one for the period 1984–1991, and one for the period 1992–2003. Table 1 presents the estimated values of b_k for both sample periods. We can see that the hump-shaped pattern in the 1984–1991 period is robust to regional adjustments. Similarly, after 1991, bank performance becomes a monotone-increasing function of size even after controlling for regional factors.[12]

Boyd and Gertler (1994) also investigate the relationship between time-average loan chargeoffs and bank size. They find that for the period

[12] We also run a regression where we allowed the coefficients b_k to vary across regions (equation 2 in Boyd and Gertler [1994]). The results were very similar.

Table 1 Size-Performance, Controlling for Regional Effects

Time Period	b_k Coefficient for Each Asset Size Class					
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
1984–1991	-0.0009 (-2.78)	0.0009 (2.36)	0.00 –	-0.0014 (-2.40)	-0.0021 (-2.52)	-0.0062 (-2.60)
1992–2003	-0.0019 (-7.90)	-0.0002 (-0.86)	0.00 –	0.0010 (1.01)	0.0018 (1.92)	0.0077 (1.37)

Notes: We use data for all insured commercial U.S. banks (except credit card banks). To construct return on assets we divide annual net income by total assets. We take time averages for each bank that existed in the base year, 1983, across the eight years in the period 1984–1991. For the period 1992–2003 we follow the same procedure using 1991 as the base year. The k size classes are the same as in Figure 4. The number of observations in the regression for the period 1984–1991 is 13,964 and for the period 1992–2003 is 11,230. The values in parentheses are t-values.

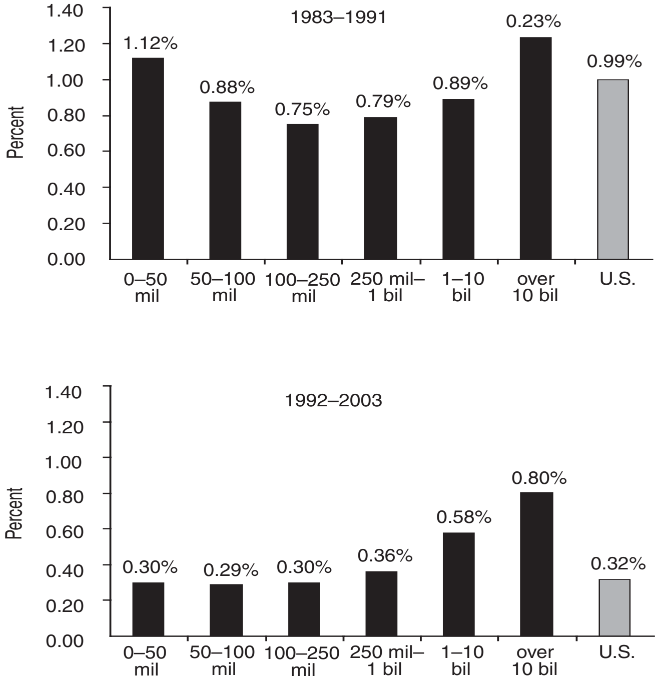
Sources: Report of Condition and Income Data (Call Report); Federal Reserve Bank of Chicago Web page.

1984–1991, the relationship has a U-shape. In other words, small and large banks tend to have higher chargeoffs to assets than medium-sized banks. This finding is taken as further evidence of the possible effects of the TBTF policy. In Figure 5 we reproduce Boyd and Gertler's result and provide the same data for the period 1992–2003. Once again, there has been a change in pattern between these two periods. For the data after 1991, the relationship between chargeoff and bank size is monotone increasing. Larger banks tend to have, on average, riskier loans.

Another variable that can be used as a proxy for bank risk is the variance of return on assets (see, for example, Berger and Mester [2003]). Boyd and Gertler (1994) do not compute this variable for their period. We provide this calculation for both subperiods in Figure 6. It is interesting to see that the variance of (annual) return on assets has significantly decreased after 1991 for all size classes. Also, the variability of return on assets does not show a monotonic relationship with the asset size of banks. In the 1984–1991 period, banks with over $10 billion in assets had a variance of return on assets that was higher than that for the previous size class (those banks with $1 to $10 billion in assets). However, this pattern is lost after 1991.

The data studied here for the period 1992–2003 are consistent with a banking system that is not necessarily distorted by the perception of potential TBTF subsidies. Under this interpretation, larger banks give riskier loans (higher chargeoffs to loans) but have a larger size of operations that allows them to better diversify those risks (lower variance on return on assets). A large

Figure 5 Chargeoffs to Loans and Size



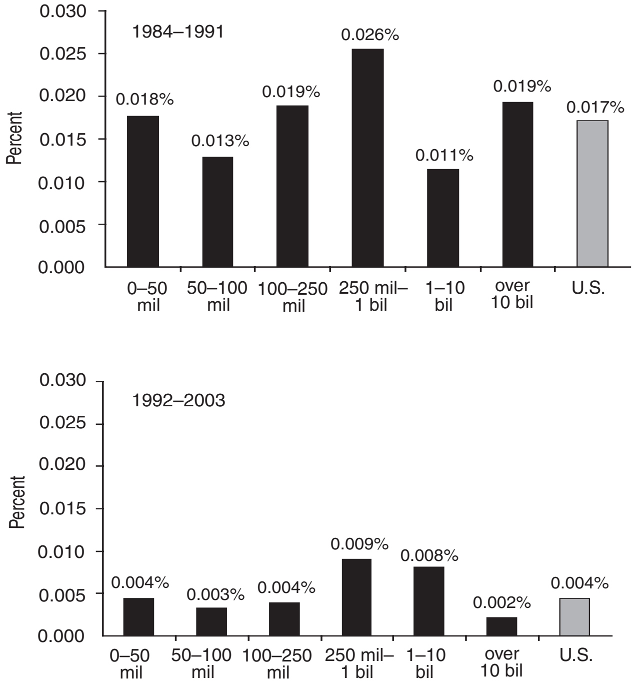
Notes: We use data for all U.S. insured commercial banks (except credit card banks). To construct chargeoffs to loans we divide annual net chargeoffs by total loans and leases. We consider each annual observation for each bank as the basic data entry in the calculation of averages across sizes (i.e., we do not take time averages for each bank). The total number of observations is around 110,000 for 1983–1991 and 120,000 for 1992–2003.

Sources: Report of Condition and Income Data (Call Report); Federal Reserve Bank of Chicago Web page.

size of operations may imply some extra cost, but the riskier loans also allow these large banks to obtain higher average returns. Ennis (2001) provides a model of banking where this kind of logic is formally studied.

At the same time, the data for 1984 to 1991 seem perhaps more consistent with the existence of a TBTF distortion. The natural question to ask then is, could it be that changes in banking regulation at the beginning of the nineties have solved the TBTF problem? The effectiveness of FDICIA in controlling TBTF has been a matter of controversy among experts. For example, Stern and Feldman (2004) argue that the post-FDICIA regime is not much different from the pre-FDICIA regime and, as a consequence, if TBTF was a problem

Figure 6 Variance of Return on Assets and Size



Notes: To construct the averages for 1984–1991, we compute the variance of annual return on assets for each bank existing in 1983 and organize the banks in size classes according to the average amount of assets they owned in that year. For the period 1992–2003 we use 1991 as the base year. We only use banks for which we have at least three annual observations. The total number of observations is around 13,000 for 1983–1991 and 10,000 for 1992–2003.

Sources: Report of Condition and Income Data (Call Report); Federal Reserve Bank of Chicago Web page.

before 1991, it is still a problem afterwards.[13] No large bank has been in trouble since the enactment of FDICIA, and it is difficult to determine the ultimate effect of the change in the regulation.

An alternative explanation for the change in the patterns observed in the data is that the late 1980s was an unusual period. The idea is that large banks

[13] According to Stern and Feldman (2004), FDICIA made explicit a set of procedures that were implicit before 1991. They judge those procedures insufficient to stop TBTF bailouts. For a more favorable view of the reforms in FDICIA, see Benston and Kaufman (1997).

specialized in certain activities (loans to less-developed countries and large commercial real-estate loans) that performed poorly during the second half of the 1980s. Boyd and Gertler (1994) discuss this interpretation but are very skeptical about its merits. They argue that medium-sized banks participated in the same set of activities as large banks but performed much better. Then, Boyd and Gertler conclude that the reason why medium-sized banks outperformed large banks is that large banks were less risk-sensitive as a consequence of the TBTF distortion.

It is interesting to note that some of the findings in this article are in accordance with the findings in the empirical literature that investigates the viability of exploiting market discipline in banking regulation. A significant portion of this literature studies the extent to which bond yield spreads reflect the financial conditions of banks. Most of this work finds that, while during the early to mid-1980s the relationship between bond yield and bank risk was weak (presumably due to implicit government guarantees), during the late 1980s and the 1990s the relationship became much stronger (see, for example, Flannery and Sorescu [1996] and the review in Flannery and Nikolova [2004]). These findings have been taken as evidence that the TBTF problem has been mitigated since the beginning of the 1990s. However, Morgan and Stiroh (2002), using data for the 1993 to 1998 period, still find that the behavior of bond spreads for those banks most likely to be subject to a TBTF policy was significantly different from that of other smaller banks and other debt-issuing corporations.

The purpose of this section was to provide some evidence to test the view that TBTF may be a latent problem in the U.S. banking system. Overall, however, it seems that looking at the data on performance across size classes does not allow any definite conclusion.

There are, of course, other ways to look for evidence of TBTF distortions. One methodology is to look at the effect of announcements about the existence of a TBTF policy over the equity value of banks. For example, O'Hara and Shaw (1990) used this strategy. They found that in September 1984, after the Comptroller of the Currency testified before Congress that certain banks were "too big to fail," the equity value of those banks increased significantly (relative to the rest of the industry).

Another way to approach the question is to study the effect of mergers on the value of the claims issued by the merging organizations. Benston, Hunter, and Wall (1995) study the prices that were bid to acquire target banks in the early to mid-1980s. They find little evidence of a TBTF-subsidy-enhancing motive in a sample of U.S. bank mergers during that period. On the other hand, Penas and Unal (2004) study changes in the return on nonconvertible bonds issued by merging banks during the 1991–1997 period. They find a significant increase in bondholder returns after a merger and that the increase is non-monotone with respect to the asset size of the bank. In particular, holders of bonds issued by mid-sized banks (especially those that after merging became relatively large within the system) are the ones that benefit the most from a merger. The authors attribute this pattern to a TBTF perception in the market for bonds.

Yet another methodology is to look at the cost-savings implications of increases in bank size. Some empirical studies have found that economies of scale exhaust at fairly modest bank sizes ($ 200 million in assets). If this is the case, then the existence of larger banks may be the consequence of a TBTF distortion. However, the empirical literature on economies of scale in banking is far from a consensus. Wheelock and Wilson (2001), for example, find that economies of scale do not exhaust until banks have at least $500 million in assets and do not find evidence of significant diseconomies of scale for larger banks (see also Hughes, Mester, and Moon 2001).[14]

3. CONCLUSION

In this article we have formally identified some basic principles that guide the behavior of banks interacting under the coverage of a government safety net, and in particular, a TBTF policy. We also studied some empirical regularities of U.S. bank performance across size classes and evaluated the extent to which they provide evidence of a significant size and risk distortion originated in a perceived TBTF subsidy.

Our conclusion is a word of caution. While, in principle, the cost of the TBTF distortions could be large, the available evidence is far from conclusive. This is an important reality to acknowledge. Several policy measures are currently being considered to reduce the potential distortions induced by TBTF (Stern and Feldman 2004). To the extent that some of these policies create new inefficiencies in the economy (by, for example, limiting the behavior of banks in particular ways), we need to be able to assess better their potential benefits. In this respect, then, it seems necessary, if not urgent, to improve our knowledge of the actual magnitude of the TBTF problem in the U.S. economy. Our reading on this matter is that the available evidence is very preliminary and in no way definitive.

[14] Assessments by credit rating agencies provide another source of useful information. Stern and Feldman (2004, Chapter 4), for example, present extensive evidence suggesting that credit rating agencies are in agreement on the existence of a TBTF policy for large banks.

REFERENCES

Benston, George J. and George G. Kaufman. 1997. "FDICIA after Five Years." *Journal of Economic Perspectives* 11 (Summer): 139–58.

______________, William C. Hunter, and Larry D. Wall. 1995. "Motivations for Bank Mergers and Acquisitions: Enhancing the Deposit Insurance Put Option versus Earnings Diversification." *Journal of Money, Credit, and Banking* 27 (August): 777–88.

Berger, Allen N. and Loretta J. Mester. 2003. "Explaining the Dramatic Changes in Performance of U.S. Banks: Technological Change, Deregulation, and Dynamic Changes in Competition." *Journal of Financial Intermediation* 12: 57–95.

Boyd, John H. and Mark Gertler. 1994. "The Role of Large Banks in the Recent U.S. Banking Crisis." Federal Reserve Bank of Minneapolis *Quarterly Review* (Winter): 319–68.

Demsetz, Rebecca S., Marc R. Saidenberg, and Philip E. Strahan. 1996. "Banks with Something to Lose: The Disciplinary Role of Franchise Value." Federal Reserve Bank of New York *Economic Policy Review* (October).

Ennis, Huberto M. 2001. "On the Size Distribution of Banks." Federal Reserve Bank of Richmond *Economic Quarterly* 87 (Fall): 1–25.

Flannery, Mark J., and Sorin M. Sorescu. 1996. "Evidence of Bank Market Discipline in Subordinated Debenture Yields: 1983–1991." *Journal of Finance* 51: 1347–77.

Flannery, Mark J., and Stansilava Nikolova. 2004. "Market Discipline of U.S. Financial Firms: Recent Evidence and Research Issues." In *Market Discipline across Countries and Industries*, ed. William C. Hunter, George G. Kaufman, Claudio Borio, and Kostas Tsatsaronis. Cambridge: The MIT Press: 87–100.

Greenspan, Alan. 2000. "Banking Evolution." Speech at the 36th Annual Conference on Bank Structure and Competition of the Federal Reserve Bank of Chicago, Chicago, Illinois.

Hetzel, Robert L. 1991. "Too Big to Fail: Origins, Consequences, and Outlook." Federal Reserve Bank of Richmond *Economic Review* 77 (November/December): 3–15.

Hughes, Joseph P., Loretta J. Mester, and Choon-Geol Moon. 2001. "Are Scale Economies in Banking Elusive or Illusive? Evidence Obtained by

Incorporating Capital Structure and Risk-Taking into Models of Bank Production." *Journal of Banking and Finance* 25 (December): 2169–208.

Keeley, Michael C. 1990. "Deposit Insurance, Risk, and Market Power in Banking." *American Economic Review* 80 (December): 1183–1200.

Morgan, Donald P., and Kevin J. Stiroh. 2002. "Too Big to Fail and Market Discipline of Banks: A Cross-Sector Test." Federal Reserve Bank of New York. Mimeo.

O'Hara, Maureen and Wayne Shaw. 1990. "Deposit Insurance and Wealth Effects: The Value of Being 'Too Big to Fail'." *Journal of Finance* 45 (December): 1587–601.

Penas, María Fabiana and Haluk Unal. 2004. "Gains in Bank Mergers: Evidence from the Bond Markets." Forthcoming *Journal of Financial Economics*.

Prescott, Edward S. 2001. "Regulating Bank Capital Structure to Control Risk." Federal Reserve Bank of Richmond *Economic Quarterly* 87 (Summer): 35–52.

Prescott, Edward S. 2002. "Can Risk-Based Deposit Insurance Premiums Control Moral Hazard?" Federal Reserve Bank of Richmond *Economic Quarterly* 88 (Spring): 87–100.

Stern, Gary H. and Ron J. Feldman. 2004. *Too Big to Fail: The Hazards of Bank Bailouts*. Brookings Institution Press. Washington, D.C.

Walter, John R. "Closing Troubled Banks: How the Process Works." Federal Reserve Bank of Richmond *Economic Quarterly* 90 (Winter): 51–68.

Wheelock, David C. and Paul W. Wilson. 2001. "New Evidence on Returns to Scale and Product Mix Among U.S. Commercial Banks." *Journal of Monetary Economics* 47: 653–74.

What Difference Would an Inflation Target Make?

Robert L. Hetzel

Numerous economists have advocated an inflation target for the United States (see, for example, Mishkin [1999] and Goodfriend [2005]). What, if anything, would an inflation target change about the way the Federal Reserve makes monetary policy? To answer this question, one must be explicit about the strategy used to achieve that target. Such a strategy might not even include inflation as an operational target. An answer to this question therefore requires specification of a policy rule—an explicit formulation of the objectives of monetary policy and the strategy for achieving those objectives.

A policy rule would clarify what, if anything, an explicit inflation target implies about other objectives. Are real output and unemployment also objectives? If a tradeoff between fluctuations of inflation around its target and fluctuations of real output around potential or trend output exists, real output and unemployment must be included as objectives along with inflation. Whether this tradeoff exists depends upon the structure of the economy. Therefore, to evaluate the prospective working of an inflation target requires explicit specification of both a policy rule and a model of the economy. Different models possess different implications for how a central bank would make an inflation target operational. How does one choose such a model?

There are two different frameworks for explaining how central banks control inflation. One framework makes an “exploitable” Phillips curve the central behavioral relationship in the control of inflation. That is, the central bank can use the real-nominal (unemployment-inflation) correlations in the empirical data as a reliable lever with which to trade off between these variables.

I gratefully acknowledge computational assistance from Elise Couper and helpful comments from Kartik Athreya, Marvin Goodfriend, Andreas Hornstein, and Alexander Wolman. Readers should not attribute the views in the paper to the Federal Reserve Bank of Richmond or to anyone else other than the author.

The other tradition, the quantity theory, makes monetary control the central behavioral relationship. That is, to control inflation, the central bank must control the rate at which nominal money grows relative to real money demand by the public. The former tradition, but not the latter, implies that the control of inflation imposes a tradeoff between variability in real output and inflation.

Two problems arise in choosing the correct model. The first is the lack of consensus over the empirical generalizations that should serve as a basis for choosing one model over the other. What lessons does one draw from the recent historical experience in the United States of rising inflation followed by disinflation? How does one summarize this experience in a way that allows a choice between competing models?

Section 1 lists four empirical and theoretical generalizations that I consider consistent with this experience and with the quantity theory framework. Section 2 discusses the consensus that exists over these generalizations. (The central bank must provide a nominal anchor.) Section 3 discusses the disagreement. (What is the nature of the Phillips curve?)

The second problem is that no exposition of a model in the quantity theory tradition exists that is useful for explaining monetary control applicable to a central bank that uses an interest rate instrument. The central insight of the quantity theory is that the nominal quantity of money can change without a prior change in real money demand. When it does, the price level must change to restore the real quantity of money demanded by the public (Friedman 1969).[1] However, existing expositions of the quantity theory assume that the central bank employs a money target. They leave unclear how a central bank, when it uses an interest rate instrument, achieves monetary control. That is, how does it avoid changes in money that produce undesired changes in prices?

Section 4 discusses monetary control when the central bank uses an interest rate instrument. Section 5 illustrates these ideas of monetary control through a model simulation that highlights periods when the central bank does and does not achieve its inflation target. The final section explains how an inflation target increases the demands placed on a central bank to communicate clearly to the public.

1. FOUR QUANTITY THEORY GENERALIZATIONS

The current monetary standard did not emerge as an explicit choice among alternative policy rules with clearly defined nominal anchors. Nevertheless, its adoption occurred as part of the debate engendered by the intellectual ferment that accompanied the 1970s stagflation. Two sets of ideas came together to

[1] Hetzel (2004a), which places the monetary policy of the Fed in a quantity theoretic framework, is complementary to this article. Goodfriend (forthcoming) and Nelson (2003) offer alternative arguments for incorporating money into models of price level determination.

provide the intellectual basis for the new standard. One set derived from the quantity theory, a tradition as old as the discipline of economics.[2] The other set, the rational expectations revolution, was new.[3] The quantity theory, with its fundamental distinction between nominal and real variables, explained the need for a nominal anchor that only a central bank could provide. Rational expectations, which emphasized the importance of policy rules in the formation of expectations, explained the importance of a credible rule. A model useful for explaining how central banks control inflation will incorporate four generalizations shaped by these two sets of ideas.

First, the price system works well to clear markets. Conceptually, a monetary economy possesses a real business cycle core that would emerge with complete flexibility of the price level (Kydland and Prescott 1982; Prescott 1986). Allowed to operate in an unhindered fashion, the price system embodied in this core allocates resources efficiently. A useful conceptual benchmark for assessing monetary policy is the difference between the values of real variables and their "natural" values represented by equilibrium in this real core.

Second, individual welfare depends upon real variables (real quantities and relative prices), not nominal magnitudes (dollar amounts). For this reason, only the central bank can give fiat money (a nominal variable) a well-defined value. The way that it does so determines the nature of the nominal anchor and the monetary standard.

An implication of the first and second generalizations is that central banks control the price level through their control of nominal money relative to real money demand. The price level is a price—the money price of goods. The assumption that the price system works means that the price level varies to clear the market for the quantity of money.[4] That is, it varies to endow the nominal quantity of money with the real quantity of money (purchasing power) desired by the public.[5] In this sense, the price level is a monetary phenomenon.

Third, individuals use information efficiently (form their expectations rationally). This generalization, plus the monetary character of the price level, implies that individuals base their expectation of the future price level on the systematic, predictable part of monetary policy (Lucas [1972] 1981). Be-

[2] Expositions of the quantity theory go back to David Hume ([1752] 1955). The most recent expositions are Friedman (1969, 1974). Friedman (1989) summarizes these earlier expositions. See also Brunner (1971).

[3] See Lucas ([1972] 1981; [1973] 1981) and Sargent ([1971] 1981).

[4] The equilibrating role played by the price level changes with fixed, rather than floating, exchange rates. In the former case, the price level varies to equilibrate the balance of payments. The nominal anchor is the foreign price level. Through the balance of payments, nominal money adjusts to provide the real money desired by the public.

[5] The real quantity of money possesses a well-defined value—the natural value associated with complete price flexibility. This value is not unique but varies with the cost of holding money (the nominal interest rate).

cause individuals' welfare depends upon their ability to distinguish real from nominal changes (an absence of money illusion), they will use the systematic part of monetary policy to forecast inflation. Of course, the central bank can render this task difficult by following a rule that yields unpredictable changes in prices.

Given that individuals base their expectations and behavior on the systematic part of the central bank's actions, it follows that achievement of an inflation target (any nominal equilibrium) requires a monetary rule—a consistent procedure by the central bank for pursuing the target (Lucas [1980] 1981). If the central bank does not target money directly, this rule must tie down the expected future value of money (the inverse of the price level, the goods price of money). Stated alternatively, in a fiat money regime, money, which is intrinsically worthless, possesses value because individuals expect it to possess value in the future. Individuals part with goods for money today only because they believe that others will accept goods for money tomorrow. The central bank must provide a nominal anchor that determines how the public forms an expectation of the future price level. To do so requires the central bank to behave in a consistent, predictable manner, that is, to follow a policy rule.

It also follows that the rule determining the price level will possess a characterization in terms of monetary control (control of a nominal variable, money) rather than control of output or unemployment (real variables). If the welfare of individuals depends upon real—not nominal—variables, and they form their expectations of inflation conformably with the monetary rule, then the central bank either cannot manipulate real variables in a systematic manner (Sargent and Wallace [1975] 1981) or will not find it desirable to do so (Goodfriend and King 1997). At issue is the interpretation of the real-nominal (unemployment-inflation) correlations summarized in Phillips curves. Can the central bank use these correlations to influence predictably the behavior of the public? Can it systematically manipulate unemployment to control inflation—or manipulate inflation to control unemployment—or trade off between the two by increasing the variability of one to reduce the variability of the other?[6]

To make the importance of a rule concrete, note that the FOMC controls only an overnight interest rate—the funds rate. A stabilizing response of the yield curve to a real shock requires a widely understood, credible rule. Consider a persistent, positive shock. Credibility for price stability means that the public assumes that the central bank will raise the funds rate by what-

[6] Milton Friedman (1968, 1977) and Robert Lucas ([1972] 1981; [1973] 1981; and 1996) argued that the real-nominal correlations summarized by Phillips curves represent a reduced form rather than a structural relationship. That is, these correlations are not invariant to the monetary rule. If the central bank attempted to "exploit" them to control unemployment, they would disappear. Similarly, they would disappear if the central bank attempted either to control inflation by manipulating unemployment or to control the joint variability of unemployment and inflation.

ever amount necessary to maintain price stability. With credibility, the entire upward shift in the yield curve will reflect higher expected future real rates of interest, not higher inflation. Conversely, with a persistent, negative real shock, credibility implies that the entire downward shift in the yield curve be real. The central bank gains nothing by attempting to offset the negative shock with inflation. Pursued as a systematic policy, the public will anticipate such behavior, and the expectation of inflation will offset the stimulative effect of the increased inflation.

The fourth and final generalization that I take as summarizing the quantity theory is that there is a fundamental difference between a relative price and the price level, which is an average of individual dollar prices. The price system decentralizes the information required to discover equilibrium relative prices (Hayek 1945) but does not do so for the price level. Price setters require some common coordinating device for changing their dollar prices that collectively allows the price level to change in a way that is compatible with the monetary policy rule and that avoids undesirable changes in relative prices. That coordinating device is a common expectation of inflation. It functions well only when the policy rule makes changes in the price level predictable.

Monetary nonneutrality arises when price setters lack a common expectational compass that provides for the coordination of the change in all dollar prices required by the policy rule.[7] Consider a policy that engenders monetary contraction requiring unpredictable reductions in the price level. Those reductions take place only through a discovery process—occurring without this coordination—by individual firms. The firm that moves first to lower its price faces the problem of strategic interaction with competitors.[8] There is a

[7] This explanation is in the misinformational spirit of Lucas ([1972] 1981) rather than the sticky-price spirit of Calvo (1983). Consider two firms, $i = 1, 2$. A firm will change $\hat{p}_i$, its dollar price, by an amount, $\hat{r}_i$, equal to the desired change in relative price, plus $\hat{p}_i^e$, the expected change in the price level:

$$\hat{p}_i = \hat{r}_i + \hat{p}_i^e. \tag{1}$$

The monetary policy rule will be consistent with some change in price level $\hat{p}$. However, if that rule is only imperfectly known to firms, they will expect a change in the price level $\hat{p}_i^e = \hat{p} + \mu_i$. The actual inflation rate $\hat{p}^m$will then differ from $\hat{p}$:

$$\hat{p}^m = \hat{p} + \omega_1 \mu_1 + \omega_2 \mu_2, \tag{2}$$

where ω_i is the expenditure share of firm i.

If the central bank announces an inflation target $\hat{p}^T$ and follows a policy rule consistent with that target, actual inflation will equal the inflation rate consistent with the rule and both will equal the inflation target ($\hat{p}^m = \hat{p} = \hat{p}^T$). In contrast, if the central bank behaves in a way that makes the inflation rate unpredictable, it will create relative price distortions that affect real variables. That is, as shown in (2), the behavior of relative prices will not wash out, but will instead affect inflation.

[8] Prices are sticky in the sense that this firm must change its price under the assumption that its competitors will not change their prices. Ball and Romer (1991) capture this strategic interaction through a multiplicity of equilibria.

positive externality to lowering one's price first that the individual firm does not capture.[9]

2. THE CENTRAL BANK SHOULD CONTROL INFLATION

A consensus now exists that the central bank determines trend inflation and must provide a nominal anchor. That consensus derives from the results of the different policy rules followed by the Fed before and after 1980. Before 1980, monetary policy did not provide a nominal anchor.[10] Policymakers based policy on the assumption that inflation was a nonmonetary phenomenon; that is, inflation possessed many sources unrelated to the degree to which central bank procedures provided for monetary control. Consequently, the policy appropriate for the control of inflation depended upon the source of the inflation. Especially, cost-push inflation was better dealt with through incomes policies than through a restriction of aggregate demand, which would raise unemployment.[11] Incomes policies ranged from occasional government interference in price setting in particular markets to full-scale wage and price controls.

Universally, policies for the control of inflation based on incomes policies failed. By default, at the end of the 1970s, governments turned to central banks exclusively to control inflation. Because central banks are the organization with a monopoly on the monetary base, that decision vindicated Friedman's hypothesis that inflation is always and everywhere a monetary phenomenon.[12] The experiment created a consensus that the central bank must provide a nominal anchor.[13]

[9] The externality comes from increasing the real quantity of money toward the amount demanded by individuals collectively.

[10] In the 1970s, through a continuing influence on expectations, the long historical experience with a commodity standard provided a nominal anchor. However, that influence disappeared by 1980.

[11] See Hetzel (1998, 2004a) for references.

[12] Friedman based his hypothesis on the empirical observation that high money growth accompanies high inflation. Once a central bank succeeds in restoring price stability, however, money loses its ability to predict prices. The apparent disappearance of money as a useful predictor obscures the validity of the Friedman hypothesis. I interpret the statement that inflation is a monetary phenomenon as follows: First, the central bank must supply the nominal anchor (provide for nominal determinacy). Second, different monetary policy rules determine different time series behavior of the price level. Third, the trend rate of inflation is under the complete control of the central bank. Note that the hypothesis does not imply that for a given rule real shocks exercise no influence on the price level.

[13] Bernanke (2005) reviewed a similar Latin American experiment. The structuralist theory of development attributed inflation to nonmonetary factors such as competition among groups for incompatibly large shares of national income. Policies based on such ideas led to high rates of inflation. Not until the end of the 1990s, when governments assigned responsibility for inflation to central banks, did Latin American countries achieve low inflation.

3. HOW DOES THE CENTRAL BANK CONTROL INFLATION?

Although there is a consensus that the central bank should control inflation, there is no consensus over how it does so. If the above quantity theory generalizations are correct, then central banks control inflation through procedures that provide for monetary control. The alternative is that they control inflation through manipulation of the unemployment rate according to a Phillips curve relationship. At issue is the nature of the Phillips curve and whether it offers a reliable lever for manipulating the behavior of the public.

Economists have suggested many possible Phillips curves.[14] Lucas ([1972] 1981) suggested the New Classical Phillips curve, and Calvo (1983) and Rotemberg (1987) suggested the New Keynesian Phillips curve. Each explain some aspects of the pre-1980 and post-1980 change in the monetary policy rule.[15]

The New Classical Phillips curve makes output fluctuations (deviations of real variables from their natural values) a function of forecast errors for inflation. Fluctuations in real variables are generated by the unpredictable component of monetary policy. Assuming that individuals form their expectations of inflation in a way that incorporates the systematic part of the monetary policy rule, the central bank cannot predictably manipulate nominal variables (inflation) to control real variables (an output gap or unemployment). Conversely, the central bank cannot manipulate these real variables to control inflation. The reason is that a systematic attempt to exploit the real-nominal correlations present in the data would make them disappear.

Stated another way, an implication of the New Classical Phillips curve is that the inflation-unemployment correlations in the data will not survive changes in the monetary policy rule (Friedman 1968; Sargent [1971] 1981; Lucas [1972] 1981, [1973] 1981). As predicted, in the 1970s, the negative correlation between unemployment and inflation disappeared in the face of sustained inflation. Furthermore, in the post-1980 period, in stabilizing inflation, monetary policy not only reduced inflation, but also maintained it without unemployment above its natural level. Moreover, both the variability of output and inflation fell. Orphanides and van Norden (2004) find parameter instability in a variety of Phillips curves estimated over the intervals 1969 to 1982 and 1984 to 2002. This evidence of structural instability runs counter to formulations of the Phillips curve that offer the policymaker a predictable tradeoff between inflation and unemployment. Examples are the original permanent trade-off formulation of Samuelson and Solow (1960) or the later NAIRU formulation of Modigliani and Papademos (1975).

[14] See McCallum (2002, footnote 38) for a long list.

[15] The former is used in flexible price rational-expectations natural-rate models (see Sargent and Wallace [1975]). The latter is used in sticky price New Keynesian models. See Goodfriend and King (1997), who use the term New Neoclassical Synthesis.

An alternative to the New Classical Phillips curve is the New Keynesian Phillips curve shown in (3).[16] Like the New Classical Phillips curve, it makes expected inflation central. Contemporaneous inflation, π_t, depends upon expected future inflation, π^e_{t+1}, and a real variable measuring the intensity of resource utilization. In a world of monopolistic competition, the latter could be the markup (price over marginal cost) relative to its profit-maximizing (natural) value (Goodfriend 2004a). Alternatively, it could be an output gap (the difference between output, y_t, and the natural or potential level of output, y^*_t). The output gap measures the extent to which price rigidities produce variations in labor supply that move output away from its natural (flexible price) value (King and Wolman 1999):

$$\pi_t = \pi^e_{t+1} + b(y_t - y^*_t). \tag{3}$$

This Phillips curve contrasts with the traditional Keynesian Phillips curve (4), where inflation depends upon lagged inflation, an output gap that measures idle (unemployed) resources, and an inflation shock, ε_t. Equation (4) captures the view that inflation shocks (changes in relative prices that pass through to the price level) initiate inflation. The lagged inflation term π_{t-1} expresses a structural persistence in inflation that exists independently of whether the monetary policy rule accommodates (propagates) the impact of relative price shocks on the price level.[17] With (4), in response to, say, a positive inflation shock ε_t, the central bank creates a negative output gap $(y_t - y^*_t)$ to limit the increase in inflation. In a model of sticky prices, a Phillips curve like (4) requires that the central bank increase output variability to reduce inflation variability:[18]

$$\pi_t = \pi_{t-1} + b(y_t - y^*_t) + \varepsilon_t. \tag{4}$$

The expectational Phillips curve (3) expresses the importance of credibility in controlling inflation. That is, in the post-1989 period, the Fed has stabilized actual inflation by stabilizing expected inflation. FOMC behavior during inflation scares provides evidence that the policy rule that provided a nominal anchor in the post-1980 period entailed a consistent effort to stabilize expected inflation at a low level (Goodfriend 1993, 2004b, and 2005; Goodfriend and King 2004; and Hetzel 2005). For example, during the 1984

[16] One reason economists turned to the New Keynesian Phillips curve is that the New Classical Phillips curve does not explain why monetary shocks impact output and employment before inflation. For a derivation of (3), see Rotemberg and Woodford (1997).

[17] That is, given the structure of the economy assumed in (3) but not (4), in empirical estimation, lagged inflation terms would appear only to that extent that the monetary rule makes them useful for predicting inflation (Sargent 1971).

[18] Based on a Phillips curve like (4), Ball (1999, Figure 3.1) and Rudebusch and Svensson (1999, Figure 5.2) present model-based estimates of a tradeoff between the standard deviation of the output gap and inflation.

inflation scare, characterized by a sharp rise in bond rates, the FOMC raised the funds rate despite falling actual inflation and a negative output gap.

Before 1980, the idea of an expectational nominal anchor would have seemed implausible. The popular consensus, formed by Keynesians and businessmen, held that powerful private sector forces (weather, the OPEC cartel, oligopolistic market structure, militant labor unions, etc.) and government deficits and regulation powered inflation. Expectations unmoored by monetary policy also drove inflation (the wage-price spiral). Only a high rate of unemployment could counter these real forces.[19]

The Volcker disinflation and its aftermath changed attitudes in two respects. The reduction of inflation from double digits to 4 percent demonstrated that central banks could control inflation. Just as important was the aftermath when monetary policy maintained moderate inflation without either sustained high unemployment or periodic recourse to high unemployment. A consensus then emerged that credibility was paramount for central bank control of inflation. A credible monetary policy can always and everywhere control inflation without resorting to high unemployment.[20]

Credibility or its absence can explain the failure of Phillips curves like (4) to explain particular episodes. Despite extreme tightness in labor markets and a positive output gap in 1999 and early 2000, inflation remained low. This episode was the converse of the early 1970s experience where inflation remained high despite an apparent negative output gap. Credibility in the first episode and in its absence in the second can explain this behavior.

Another kind of evidence in favor of an expectational Phillips curve like (3) is the failure of relative price shocks to exercise a persistent influence on inflation when the central bank possesses credibility.[21] Relative price shocks do occasionally pass through to the price level.[22] With credibility, however, a positive, relative price shock, for example, that passes through to the price

[19] This view still appears in the Phillips curve (4), where inflation shocks initiate inflation and intractable inflation persistence propagates them. To maintain expected inflation equal to its target, the central bank must raise unemployment to counter these shocks.

[20] After 1979, the Fed abandoned its former policy rule without committing to a new one in a credible, public way. The public had to learn the new rule over time. While this learning occurred, expectations were not "rational" in the sense of being consistent with the new rule. The Fed had to incur costs to establish credibility. During inflation scares, it had to shock the economy through unanticipated monetary contractions. The real-nominal tradeoffs of the Phillips curve occur in establishing credibility but not in controlling inflation with credibility.

[21] Consider oil price shocks. In early 1999, the price of a barrel of crude oil was around $10. In March 2005, it reached $55. The unemployment rate, which fell from July 2003 onward, did not provide an offset to this "inflation shock" in the last part of this period. As indicated by the behavior of bond rates, financial markets have not expected this increase in the relative price of oil to increase inflation other than transitorily.

[22] If the central bank had to respond directly to such price level changes, it could adversely affect unemployment. The Friedman (1960) long-and-variable lag criticism of price level targeting would apply. If the central bank maintains expected inflation equal to its target and then moves the funds rate so that the real rate tracks the natural rate, money creation follows changes in real money demand. Money is not an independent influence (see Section 4). If the central bank

level does not create an expectation of further inflation. Consequently, there is no expectational coordinating mechanism to propagate a general price rise (note generalization (4) of Section 1). As a consequence of central bank use of an interest rate target as opposed to a money target, there is no real balance effect to reverse the contemporaneous price rise, and the price level can drift. However, over time, such changes tend to wash out.[23]

Unlike the Lucas Phillips curve of the New Classical model (2), the Calvo-Rotemberg Phillips curve of the New Keynesian model (3), which assumes sticky (infrequently changed) prices, does offer a predictable tradeoff between real and nominal variables. However, even though the central bank can systematically exploit a tradeoff between real and nominal variables, it should not. A policy of price stability is welfare maximizing because it avoids the relative price distortions that inflation causes with intermittent price setting.[24] With either type of Phillips curve, central bank control of inflation through manipulation of the real-nominal correlations of the Phillips curve is not an option, either because it is infeasible or because it is undesirable.

4. MONETARY CONTROL WITH AN INTEREST RATE INSTRUMENT

If central bank control of inflation does not possess a characterization in terms of manipulation of a Phillips curve relationship, it must possess a characterization in terms of monetary control. However, with central bank use of an interest rate instrument, monetary control does not require a target for money.

An understanding of monetary control with an interest rate instrument begins with the fact that the market interest rate comprises a nominal and a real component. For each component, there is a unique value consistent with maintaining inflation equal to target. Monetary policy must set the first equal to its inflation target and discover the value of the second to make it equal the natural interest rate (the real interest rate that would prevail with complete price flexibility).

created and destroyed money to offset transitory movements in inflation, money creation would become an independent influence with unpredictable results.

[23] The forecast error associated with a prediction of the future price level can still grow very large as the forecast horizon lengthens. If the central bank wanted to assure price stability as opposed to inflation stability, it would need to establish credibility for a negative correlation between actual inflation and expected inflation rather than just a zero correlation.

[24] See Goodfriend and King (1997, 2001) and Wolman (2001). Rotemberg and Woodford (1999, 74) state: "[E]ven though our proposed welfare criterion ... assigns ultimate importance only to the efficiency of the level of real activity..., it in fact justifies giving complete priority to inflation stabilization as opposed to output stabilization."

King and Wolman (1999, 350) state: "[T]he monetary authority should ... make the price level the sole objective of monetary policy .... [P]rice level stabilization policy is optimal in a very specific sense: it maximizes utility of the representative individual."

The central bank must stabilize the public's expectation of (trend) inflation by following a rule that keeps that expectation equal to its inflation target. Consider an inflation scare (Goodfriend 1993). To maintain expectational credibility (the nominal anchor), the central bank must be willing to create an unexpected negative output gap by imparting a monetary shock.[25]

Given equality between expected inflation and the central bank's inflation target, the central bank can take for granted stability in the inflation premium in the interest rate. It can then vary its interest rate instrument to produce predictable changes in the real rate. It can concentrate on the sole objective of making the real interest rate track the natural interest rate. In doing so, it allows the price system to maintain resource utilization at its natural level (keep $y_t - y_t^*$ at zero). Real output grows in line with potential output.

Expected inflation then drives both actual inflation and money growth beyond changes in real money demand, which the central bank accommodates as a consequence of maintaining an interest rate peg. In a sense, there is a Friedman k-percent rule for money growth, which equals the k-percent inflation target plus whatever additional amount is required to accommodate changes in real money demand. However, changes in nominal money rather than changes in the price level provide the public with desired changes in real money. The central bank provides for desired changes in real money (consistent with the behavior of natural output) through changes in nominal money, thereby obviating the need for price level changes (beyond those compatible with the inflation target).

A policy rule that works poorly to maintain equality between the real and natural interest rate engenders excess money creation (destruction). With a policy rule that permits base drift in the price level, these monetary emissions (absorptions) force changes in the price level. They are the real-world counterparts to the helicopter drop of money used by Friedman (1969) in expositions of the quantity theory. Now, changes in the price level provide the public with desired changes in real money. Although a central bank with an interest rate instrument does not target money directly, to stabilize the price level it must possess a rule that provides for monetary control in the sense that money creation only accommodates prior changes in real money demand rather than forcing changes in the price level.

One can use this definition of monetary control to understand FOMC procedures. Because the FOMC does not possess observations of the natural interest rate, it requires an indicator that registers misalignment between the real and natural rate. Hetzel (2004a, 2005) argues that the FOMC uses a growth gap indicator. It raises the funds rate when the economy is growing faster than its estimated potential growth, and conversely.

[25] An ability to affect real variables through a monetary shock is different from controlling real variables in a systematic manner.

From a different but equivalent perspective, Broaddus and Goodfriend (2004) argue that FOMC procedures prevent emergence of a markup gap—the difference between the markup and its natural or profit maximizing value.[26] When the gap falls, the FOMC raises the (real and nominal) funds rate to restrain aggregate demand and restore the optimal output gap, and vice versa. Because the FOMC does not observe the output gap, it requires an indicator. The indicator is the degree of stress on resource utilization, which the FOMC synthesizes from extensive review of economic statistics.

5. A REAL-WORLD HELICOPTER DROP OF MONEY

The language of economics is the language of tradeoffs. To use this language to understand the implications of an inflation target requires a model. I use the New Keynesian model to illustrate the control of inflation given that the central bank does not find it desirable to trade off inflation variability against output variability. The model places the control of inflation in the context of monetary control rather than manipulation of the real-nominal relationship of a Phillips curve. (The appendix exposits the model.)

The intent of the simulation is to provide a real-world counterpart to the exogenous increase in money assumed in expositions of the quantity theory. In the simulation, a persistent productivity shock raises the natural interest rate. Interest rate smoothing by the central bank causes the money stock to rise. Although money is endogenous, the price level must rise to maintain real money equal to the amount the public demands.[27]

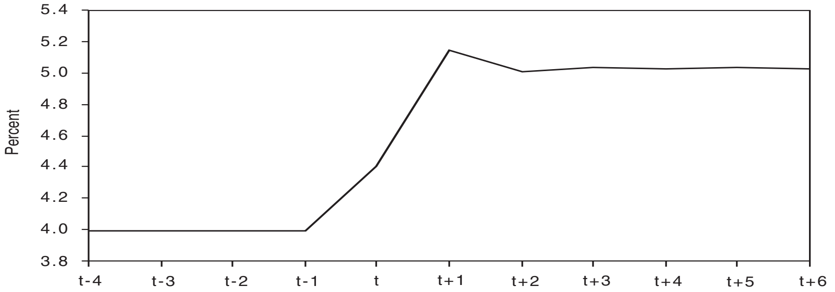
With the policy rule (5), the nominal anchor is an inflation target. The central bank responds to the difference between actual (π) and targeted inflation (π^T). Changes in the interest rate (R_t) that the central bank sets exhibit inertia relative to a base value R^*_{t-1}, which is the average of the prior period's actual interest rate and the interest rate in the absence of shocks (the steady state interest rate), $\bar{R}_{t-1}$; that is, $R^*_{t-1} = \frac{1}{2}\left[R_{t-1} + \bar{R}_{t-1}\right]$:

$$R_t = R^*_{t-1} + \frac{1}{s}\left(\pi_t - \pi^T\right). \tag{5}$$

[26] In a world of monopolistic competition, the markup is the difference between price and marginal cost. Firms raise prices when they expect the markup to remain persistently below its profit maximizing value.

[27] In his exposition of the quantity theory, Friedman (1969, 4) examines the consequences of the following conceptual event: "[O]ne day a helicopter flies over this community and drops an additional $1,000 in bills from the sky." In the simulation, the helicopter drop comes from interest rate smoothing by the central bank at the time of a real disturbance that raises the natural interest rate.

Figure 1 Interest Rate



Because of this interest rate smoothing, the interest rate set by the central bank responds only with a lag to changes in the natural interest rate. That lag causes money creation and fluctuations in inflation around the inflation target. The point is that inflation control requires both a credible inflation target and procedures that vary the interest rate instrument so that it tracks the natural interest rate.

By definition, a central bank is the organization with a monopoly on the creation of the monetary base (money in the model). Consequently, a relationship will exist between the monetary base and the central bank's policy instrument, the interest rate. Equation (6) specifies the relationship consistent with (5):

$$\Delta ln\left(M_t\right) = \pi^T + \Delta lnc_t + v\Delta lnX_t + s\left(R_t - R^*_{t-1}\right). \tag{6}$$

The inflation target π^T determines expected and actual trend inflation. Money growth increases one for one with increases in π^T. (Δln is the change in the natural logarithm.) The interest rate peg causes changes in nominal money to move with changes in the public's demand for real money. The two terms Δlnc_t and $v\Delta lnX_t$ capture the effect of changes in real money demand on nominal money. Under the assumption of an elasticity of demand for real money with respect to real consumption (c_t) of one, a 1 percent change in real consumption produces a 1 percent change in nominal money. The $v\Delta lnX_t$ term captures the effect on nominal money of changes in the opportunity cost of holding money.[28] The last term of (6), $s\left(R_t - R^*_{t-1}\right)$, relates the interest rate smoothing of the central bank to its money creation.[29]

The real disturbance is a technology shock that raises trend output (consumption) growth from zero to 1 percent in period t.[30] The natural (steady state real) interest rate rises commensurately with the growth rate of consumption. The price of contemporaneous consumption, the real interest rate, must rise to reconcile the consumer to a pattern of consumption that now favors the future. With an assumed rate of time preference β equal to 0.97125, before the increase in productivity growth, the real interest rate is 3 percent. With an inflation target π^T of 1 percent, the nominal interest rate is 4 percent. After the productivity shock raises the trend growth rate of consumption to 1 percent, the trend nominal interest rate rises to 5 percent.

Figure 1 displays the behavior of the interest rate. In the initial period of the productivity increase, the central bank limits the interest rate increase by pulling the interest rate toward a base value (the average of the prior period's interest rate and the prior period's steady state interest rate). In period t, through money creation, the central bank keeps the interest rate at about 4.4 percent. A one-time increase in the nominal money stock of 1 percent occurs about equally in periods t and $t + 1$ (Figures 2 and 3).

The increase in consumption increases the demand for real money while the increase in the nominal interest rate decreases it. As a consequence of its interest rate target, the central bank accommodates the net change in demand. The situation is different for the money the central bank creates through its interest rate smoothing. That money creation emerges as a consequence of the central bank's effort to resist the rise in the interest rate. The public must adjust to it. Now, portfolio balance by the public requires a rise in the price level commensurate with the rise in money. With no change in the price level, there is an excess supply of money and an excess demand for bonds.[31]

The jump in real consumption in period t makes consumption in t high relative to consumption in $t + 1$ (Figure 4). In t, despite the rise in trend

[28] $X_t = \frac{R_t}{1+R_t} \cdot \frac{c_t}{w_t}$. X_t measures percentage changes in real money demand from (22). Because c_t and w_t (the real wage) move together leaving the ratio unchanged, the X_t term measures the change in the demand for money arising from changes in the interest rate, R_t.

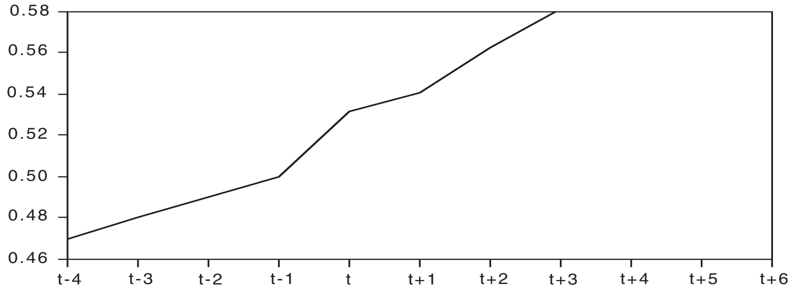
The elasticity of the demand for real money with respect to the interest rate, υ, is set at -0.038. The number is the estimate from Porter and Small (1989, 247) of the long-run elasticity of money demand. They assume that the effect of a change in the opportunity cost of holding money requires six quarters to work itself out fully. Because I assume that a period is one year, the number used here is two-thirds of the number they estimate.

[29] In the simulation, the interest rate smoothing parameter (s) equals 1.3.

[30] In (26), a_t increases by one percentage point.

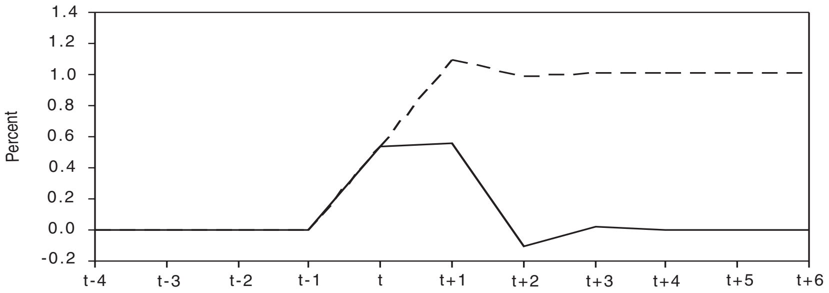
[31] To see the need for portfolio rebalancing, consider the marginal return and cost for holding money. The nominal interest rate is the cost for holding money. Equating the far left and far right sides of (15) and rearranging yields (7), which expresses the interest rate as a price that relates flows to stocks (the ratio of the flow of liquidity services from an additional dollar to the discounted marginal value of an additional future dollar of resources):

Figure 2 Money



Notes: Natural log of money.

Figure 3 Change in Money Due to Rate Smoothing

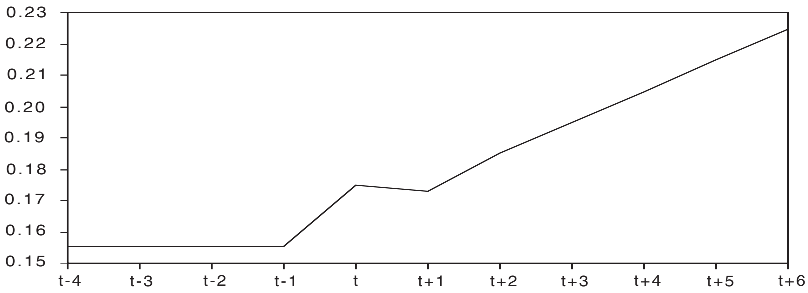


Notes: The solid line shows the changes in each period due to interest rate smoothing. The dashed line shows the cumulative changes.

$$-\frac{\mu_t h'(\cdot)\left(\frac{1}{P_t c_t}\right)}{\beta E_t \frac{\lambda_{t+1}}{P_{t+1}}} = R_t. \tag{7}$$

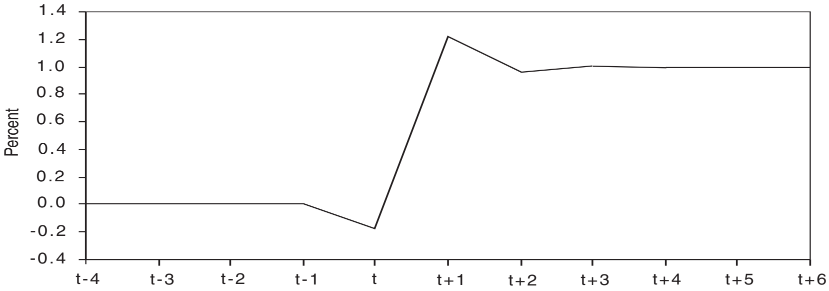
The money creation from interest rate smoothing reduces the marginal benefit from holding money. The marginal value of real money—that is, the magnitude of $h'(m_t c_t)$—declines. The return to money then falls short of the return on bonds, R_t. The individual rebalances his/her portfolio by attempting to move from money to bonds. Without a rise in the price level, this rebalancing raises the yield on bonds. The fall in the bond yield stimulates consumption. Because of price stickiness, increased nominal demand translates into increased real consumption (Figure 4).

Figure 4 Consumption



Notes: Natural log of consumption.

Figure 5 Expected Consumption Growth



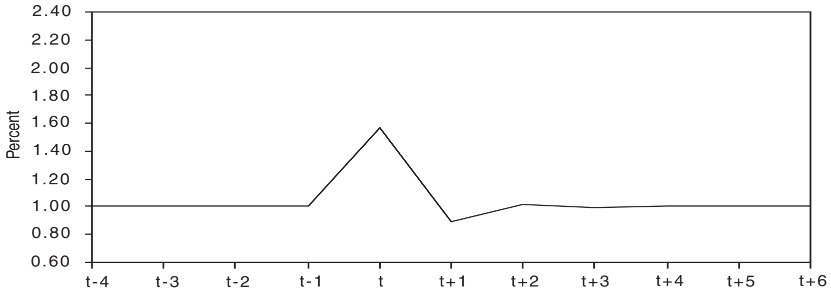
Notes: This series shows the expected percent change in consumption, $ln(C_{t+1}) - ln(C_t)$.

productivity growth, expected consumption growth is therefore negative (Figure 5), which temporarily restrains the rise in the interest rate.

Offsetting this negative influence is a rise in expected inflation above its trend level of 1 percent. In t, the public expects in $t + 1$ an additional rise in prices of 0.6 percent due to the catch-up price rises from firms whose price setting was constrained in t (Figure 6). On net, in t, the interest rate rises about 0.4 percent.

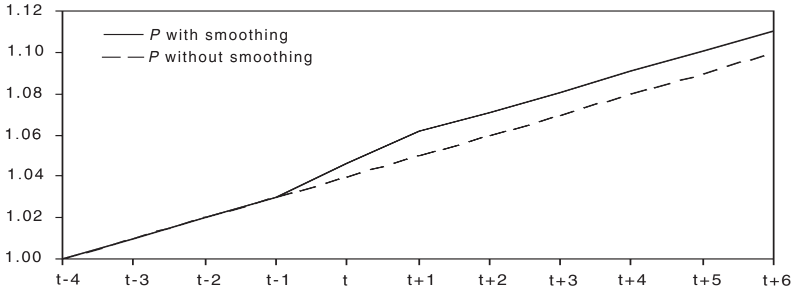
As shown by Figure 7, the money creation that arises out of interest rate smoothing causes the price level to rise by 1 percent relative to trend. The increase in money per unit of consumption is less because of the fall in the demand for real money due to the increase in the bond rate and the accommo-

Figure 6 Expected Inflation



Notes: This series shows the expected inflation, $ln(P_{t+1}) - ln(P_t)$.

Figure 7 Price Level With and Without Interest Rate Smoothing

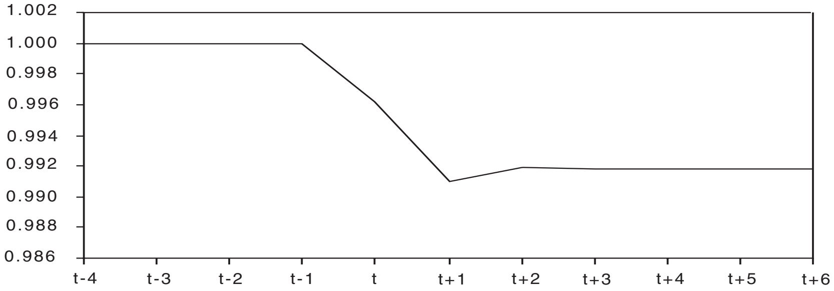


Notes: Natural log of price level, normalized so that $P_{t-4} = 1$. The solid line shows the price level with interest rate smoothing. The dashed line shows it without interest rate smoothing.

dation of this fall in demand due to the central bank's rate peg. Figure 8 shows a fall in the ratio of the real money the public desires relative to consumption, $[M_t/(P_t c_t)]$. Figure 9 shows the ratio of money per unit of consumption and the price level. The 1 percent rise in the price level (relative to trend) requires only a 0.2 percent rise in money per unit of consumption because of the 0.8 percent fall in the demand for money per unit of consumption.[32]

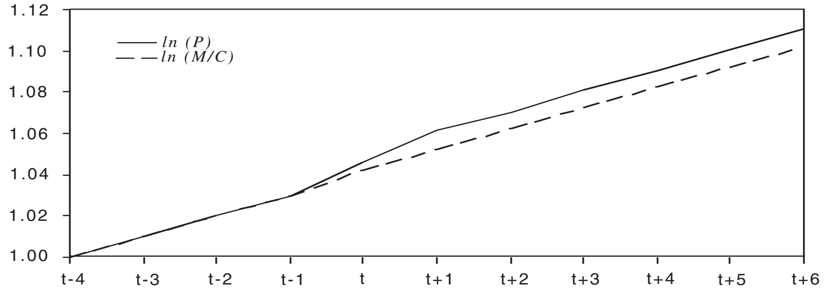
[32] Note that the individual price setters cannot index the change in their dollar prices to the money stock as protection against monetary disturbances. The relationship between money and

Figure 8 Money Demand



Notes: Natural log of inverse of velocity, $ln(m/(CP))$, normalized so that $ln(m/(CP))_{t-4}$ equals 1.

Figure 9 Money per Unit of Consumption and the Price Level



Notes: The solid line is the log of the price level; the dashed line is $ln(M/C)$. Series normalized so that $ln(Pt - 4)$ and $ln((M/C)t - 4)$ equal 1.

6. WHAT DIFFERENCE WOULD AN INFLATION TARGET MAKE?

Goodfriend (2005) argues that an inflation target need not change the way that the Fed operates because it already uses an implicit inflation target. His argument, however, does not address concerns of critics of an explicit target. They argue that such explicitness might cause the Fed to abandon its dual mandate

prices is a complicated one that depends upon the behavior of the interest rate and real money demand.

by reducing inflation variability by increasing output variability. However, I argue here that the concern is misplaced. The control of inflation requires monetary control rather than manipulation of output and unemployment in a way constrained by the tradeoffs apparently offered by a Phillips curve.

An inflation target would require better communication with the public. Like all institutions in a constitutional democracy, the long-run viability of the Fed as an institution depends upon public support, which derives from public understanding of its objectives and the procedures for achieving those objectives. An inflation target possesses the potential for enhancing that understanding. However, the standard for communication becomes more demanding than it would in a world without explicit objectives. At present, justification of changes in the funds rate can rely on a commonsense appeal to the contemporaneous behavior of the economy. In contrast, with an inflation target, justification of changes in the funds rate will derive from a need to achieve the inflation target.

The difficulty of communication arises because the funds rate may need to change in a counterintuitive way, given the behavior of inflation and the inflation target. The Fed will need a model to explain the relationship between funds rate changes and achievement of its inflation objective. For example, the funds rate may need to change continually (to track the natural interest rate) even though inflation stays on target. In this event, the public will see a positive correlation between strength in real economic activity and the funds rate, but no correlation between inflation and the funds rate. Despite appearances, monetary policy is controlling inflation, not real growth.

Adoption of an inflation target is just one step on a longer journey of making monetary policy procedures explicit. An inflation target is only the starting point for full communication with the public. Clarification of its implications for monetary policy requires specification of the policy rule that supports it. Clarification of its implications for the tradeoffs required to achieve it requires specification of the structure of the economy (a model). Although the required communication will be intense and challenging, it will advance the communication necessary for any institution that is part of U. S. constitutional democracy.

APPENDIX: THE MODEL

The model is from Wolman (1998).[33] Equation (8) represents consumers' preferences. The consumption aggregate is c_t; leisure, l_t; the discount rate, β; and a parameter measuring the value of leisure, χ:[34]

$$E_t = \sum_{t=0}^{\infty} \beta^t \cdot [ln\,(c_t) + \chi \cdot l_t]\,. \tag{8}$$

The consumer's budget constraint is

$$c_t + \frac{M_t}{P_t} + \frac{B_t/P_t}{1+R_t} = \frac{M_{t-1}}{P_t} + \frac{B_{t-1}}{P_t} + w_t n_t + d_t + \frac{S_t}{P_t}, \tag{9}$$

where the price level is P_t; the nominal money the consumer carries over into $t+1$ is M_t; the quantity of one-period nominal zero-coupon bonds that mature in $t+1$ is B_t; the market interest rate on these bonds, R_t; the real wage, w_t; work time, n_t; real dividend payments by firms, d_t; and lump sum transfers of money from the central bank, S_t.

The time constraint is

$$n_t + l_t + h\,[M_t/(P_t c_t)] = E, \tag{10}$$

where E is the time endowment. Transactions time, $h\,[M_t/(P_t c_t)]$, varies inversely with real money balances (liquidity services) measured as the fraction of expenditures the consumer holds as money balances. With $M_t/P_t = m_t$, transactions time is $h(m_t/c_t)$, $\partial h/\partial c > 0$, and $\partial h/\partial m < 0$.

1. CONSUMER CHOICE AND THE DEMAND FOR MONEY

The individual maximizes utility by choosing c_t, l_t, n_t, B_t, and M_t to maximize (8) subject to (9) and (10). The Lagrange multipliers on the latter two budget constraints are, respectively, λ_t (the marginal utility value of an extra unit of goods) and μ_t (the marginal utility value of time). The first order conditions are as follows:

[33] For further exposition, see Wolman (1997, 1999, and 2001). For a definition of equilibrium, see Wolman (1998). Goodfriend (2004a) provides an heuristic overview.

[34] $\chi = 0.77$ (Wolman 1998).

$$\frac{1}{c_t} = \lambda_t - \mu_t \cdot h'(\cdot)\left(\frac{m_t}{c_t^2}\right), \tag{11}$$

$$\chi = \mu_t, \tag{12}$$

$$\mu_t = w_t \cdot \lambda_t, \tag{13}$$

$$\frac{\lambda_t}{P_t} = \beta \cdot (1 + R_t) \cdot E_t \frac{\lambda_{t+1}}{P_{t+1}}, \text{ and} \tag{14}$$

$$-\frac{\mu_t}{P_t} \cdot h'(\cdot)\left(\frac{1}{c_t}\right) = \frac{\lambda_t}{P_t} - \beta E_t \frac{\lambda_{t+1}}{P_{t+1}} = \beta E_t \frac{\lambda_{t+1}}{P_{t+1}} R_t. \tag{15}$$

According to (11), the individual equates the marginal value of consumption, $\frac{1}{c_t}$, with the marginal value of real resources, λ_t, and the marginal value of the time foregone from that consumption, $\mu_t \cdot h'(\cdot)\left(\frac{m_t}{c_t^2}\right)$. The marginal value of time comes from (12). From (13), the individual allocates time between labor and leisure to equate the marginal rate of substitution between goods and leisure, μ_t/λ_t, to the wage rate, w_t. In (14), $\frac{\lambda_t}{P_t}$ measures the marginal utility value of an additional dollar of goods. (A doubling of the price level halves the real value of a dollar of goods.) The gross rate of interest, $(1 + R_t)$, equates the marginal value of a dollar today with the discounted expected marginal value of a dollar tomorrow.

The first order condition (15) expresses the equality between the marginal benefits and costs of holding money.[35] The utility value of the time gained from holding an additional dollar equals $-\frac{\mu_t}{P_t} \cdot h'(\cdot)\left(\frac{1}{c_t}\right)$ with $h'(\cdot) < 0$. The marginal cost of holding an additional dollar equals the marginal value of a dollar of goods today minus the discounted value of the expected gain from having an additional dollar of goods tomorrow, $\frac{\lambda_t}{P_t} - \beta E_t \frac{\lambda_{t+1}}{P_{t+1}}$. From (14), the latter equals the discounted expected future nominal value of the interest paid on bonds, $\beta E_t \frac{\lambda_{t+1}}{P_{t+1}} R_t$. The reason for the discount factor β is that the marginal benefit of holding an additional dollar is measured for the current period and the marginal cost, R_t, for the future period.

One can also make the measurement of marginal benefit and cost comparable by comparing the return from investing to the marginal cost.[36] The gain

[35] McCallum (1983) and McCallum and Goodfriend (1987) derive a money demand function within the above optimizing framework.

[36] The remainder of this section is from Wolman (1997, 6).

in transactions time from holding an additional dollar is $-h'(m_t/c_t) \cdot \frac{1}{P_t c_t}$. The value of an additional unit of time spent working when invested in a bond is $P_t \cdot w_t \cdot (1 + R_t)$. The return to holding an additional dollar, therefore, is $-P_t \cdot w_t \cdot (1 + R_t) \cdot h'(m_t/c_t) \cdot \frac{1}{P_t c_t}$. On the other hand, the cost of holding an additional dollar is the interest forgone, R_t. Equating the marginal benefit and cost of liquidity services, after rearrangement, yields

$$-h'(m_t/c_t) = \frac{R_t}{1 + R_t} \cdot \frac{c_t}{w_t}. \tag{16}$$

Equation (17) is a particular functional form for $h(\cdot)$ that relates real money balances inversely to transactions time:

$$h(m_t, c_t) = \kappa \cdot (m_t/c_t)^{\frac{-1}{\gamma}}, \gamma \in (0, 1). \tag{17}$$

Using (17) in (16) yields a money demand function,

$$\frac{\kappa}{\gamma} \cdot (m_t/c_t)^{-1-1/\gamma} = \frac{R_t}{1 + R_t} \cdot \frac{c_t}{w_t}, \gamma \in (0, 1). \tag{18}$$

With (18), real money becomes infinite as the interest rate goes to zero. A liquidity trap arises in that there is no point at which the public becomes satiated with real money. However, a liquidity trap has never been observed. For the case of Japan, see Hetzel (2003, 2004b) and, for the United States, see Wolman (1997). Satiation is the empirically relevant case.[37]

In order to attain satiation, shopping time must cease falling at some finite level of real money balances. This phenomenon cannot occur with (17) because the function asymptotes to zero. However, adding a constant to $h'(\cdot)$ makes $h(\cdot)$ cease falling and turn up at some point. At this point, satiation occurs. That is, the public has no reason to hold additional real money balances. Wolman (1997) reformulates (17) in this fashion: $h'(m_t/c_t) = \phi - (\kappa/\gamma) \cdot (m_t/c_t)^{-1-1/\gamma}, \phi \geq 0$. With $\nu \equiv -\gamma/(1 + \gamma)$ and $A \equiv (\kappa/\gamma)^{-\gamma/(1+\gamma)}$, then

$$h'(m_t/c_t) = \phi - A^{-1/\nu} \cdot (m_t/c_t)^{1/\nu} \text{ with } \nu < 0, A > 0. \tag{19}$$

The transactions time function then becomes

[37] The issue is important in dealing with the zero bound problem. If the market interest rate falls to zero, the central bank can switch to a reserve aggregate target (Goodfriend 2000; Hetzel 2003, 2004b). Money creation then exerts its influence on nominal expenditure through portfolio rebalancing. One can see the power of this effect in the historical data during periods when monetary policy procedures did not provide for monetary control. Nominal and real expenditure followed money creation, on average, with a two-quarter lag (Friedman and Schwartz 1963; Friedman 1989).

$$h\left(m_t/c_t\right) = \phi \cdot \left(m_t/c_t\right) - \frac{v}{1+v} A^{-1/v} \cdot \left(m_t/c_t\right)^{\frac{1+v}{v}} + \Omega \tag{20}$$

when $m_t/c_t < A \cdot \phi^v$ and $h\left(m_t/c_t\right) = h\left(A\phi^v\right) = \Omega + \frac{1}{1+v} A\phi^{1+v}$ when $m_t/c_t \geq A \cdot \phi^v$ with Ω a nonnegative constant equal to the minimum amount of shopping time. Shopping time decreases until real money balances reach $A \cdot \phi^v$ and then remains unchanged.

With (21), the expression for the equality of the marginal cost and benefit of holding money becomes

$$-\phi + A^{-1/v} \cdot \left(m_t/c_t\right)^{1/v} = \frac{R_t}{1+R_t} \cdot \frac{c_t}{w_t}. \tag{21}$$

The public's demand for money function then becomes[38]

$$m_t/c_t = A \cdot \left[\frac{R_t}{1+R_t} \cdot \frac{c_t}{w_t} + \phi\right]^v. \tag{22}$$

Sticky Prices

Aggregate consumption is a weighted average of different goods, $c_t = [\int c(w)^{\frac{\varepsilon-1}{\varepsilon}} d\omega]^{\frac{\varepsilon}{\varepsilon-1}}$. Firms divide into two groups, which set their product prices either in odd numbered or even numbered periods.[39] Because all firms face demand curves with constant elasticity ε, aggregate consumption equals

$$c_t = c(c_{0,t}, c_{1,t}) = \left(\frac{1}{2} \cdot c_{0,t}^{\frac{\varepsilon-1}{\varepsilon}} + \frac{1}{2} \cdot c_{1,t}^{\frac{\varepsilon-1}{\varepsilon}}\right)^{\frac{\varepsilon}{\varepsilon-1}}, \tag{23}$$

where $c_{o,t}$ and $c_{1,t}$ represent consumption in period t of goods with prices set, respectively, in the current and previous period.

The demand for each good equals

$$c_{j,t} = \left(\frac{P_{t-j}^*}{P_t}\right)^{-\varepsilon} \cdot c_t \text{ with } j = 0, 1. \tag{24}$$

P_{t-j}^* is the time t price in dollars of the good with price set in period $t - j$. The time t price level is

$$P_t = \left[\frac{1}{2} \cdot (P_t^*)^{1-\varepsilon} + \frac{1}{2} \cdot (P_{t-1}^*)^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}. \tag{25}$$

The firms' production functions are

[38] The following parameter values are from Wolman (1998): $A = 0.017$ and $\phi = 0.0014$.

[39] $\varepsilon = 10$ (Wolman 1998).

$$c_{j,t} = a_t \cdot n_{j,t}, \tag{26}$$

where $n_{j,t}$ is the amount of labor employed in period t by a firm that set its price in period $t - j$. Labor productivity is a_t. Firms' real profits are

$$\frac{P^*_{t-j}}{P_t} \cdot c_{j,t} - w_t \cdot n_{j,t}. \tag{27}$$

The firms that are free to set prices in period t set a relative price $\frac{P^*_t}{P_t}$ that maximizes the present discounted value of expected profits (28) derived from (27) using (24) and (26):

$$\begin{aligned} &c_t \cdot \left[\left(\frac{P^*_t}{P_t} \right)^{1-\varepsilon} - \frac{w_t}{a_t} \cdot \left(\frac{P^*_t}{P_t} \right)^{-\varepsilon} \right] + \\ &\beta E_t \frac{\lambda_{t+1}}{\lambda_t} \cdot c_{t+1} \cdot \left(\frac{P^*_t}{P_{t+1}} \right)^{1-\varepsilon} - \frac{w_{t+1}}{a_{t+1}} \cdot \left(\frac{P^*_t}{P_{t+1}} \right)^{-\varepsilon}. \end{aligned} \tag{28}$$

The optimal price comes from differentiating (28) with respect to P^*_t, setting the result equal to zero, and solving for P^*_t:[40]

$$P^*_t = \frac{\varepsilon}{\varepsilon - 1} \cdot E_t \left(\rho \frac{P_t w_t}{a_t} + (1 - \rho_t) \frac{P_{t+1} w_{t+1}}{a_{t+1}} \right) \tag{29}$$

with

$$\rho_t \equiv \frac{\lambda_t c_t}{\lambda_t c_t + \beta \lambda_{t+1} c_{t+1} \left(\frac{P_{t+1}}{P_t} \right)^{\varepsilon - 1}}.$$

The firm sets its dollar price as a constant markup over the present discounted value of a weighted average of the nominal marginal cost in the two periods for which the price is fixed. The weights on marginal cost in the two periods are given by the fraction of marginal revenue contributed in the particular period.

REFERENCES

Ball, Laurence, and David Romer. 1991. "Sticky Prices as Coordination Failure." *The American Economic Review* 81 (June): 539–52.

[40] For a derivation of (3) from a relationship like (28), see Rotemberg and Woodford (1997) and Woodford (2003).

Bernanke, Ben S. 2005. “Inflation in Latin America: A New Era?” Remarks at the Stanford Institute for Economic Policy Research. Economic Summit, Stanford, Calif. February 11.

Broaddus, J. Alfred, Jr., and Marvin Goodfriend. 2004. “Sustaining Price Stability.” Federal Reserve Bank of Richmond *Economic Quarterly* 90 (Summer): 3–20.

Brunner, Karl. 1971. “A Survey of Selected Issues in Monetary Policy.” *Schweizerische Zeitschrift für Volkswirtschaft und Statistik*. 107 (March): 1–146.

Calvo, Guillermo A. 1983. “Staggered Prices in a Utility Maximizing Framework.” *Journal of Monetary Economics* 12 (September): 383–98.

Friedman, Milton. 1960. *A Program for Monetary Stability*. New York: Fordham University Press.

______________. 1968. “The Role of Monetary Policy.” *American Economic Review* 58 (March): 1–17.

______________. 1969. The Optimum Quantity of Money. In *The Optimum Quantity of Money*, ed. Milton Friedman. Chicago: Aldine: 95–110.

______________. 1974. A Theoretical Framework for Monetary Analysis. In *Milton Friedman’s Monetary Framework: A Debate with His Critics*, ed. Robert J. Gordon. Chicago: The University of Chicago Press.

______________. 1977. “Nobel Lecture: Inflation and Unemployment.” *Journal of Political Economy* 85 (June): 451–72.

______________. 1989. The Quantity Theory of Money. In *The New Palgrave Money,* ed. John Eatwell, Murray Milgate, and Peter Newman. New York: W. W. Norton: 1–40.

______________, and Anna J. Schwartz. 1963. “Money and Business Cycles.” *Review of Economics and Statistics* 45 (February): 32–64.

Goodfriend, Marvin. 1987. “Interest Rate Smoothing and Price Level Trend-Stationarity.” *Journal of Monetary Economics* 19 (May): 335–48.

______________. 1993. “Interest Rate Policy and the Inflation Scare Problem.” Federal Reserve Bank of Richmond *Economic Quarterly* 79 (Winter): 1–24.

______________. 2000. “Overcoming the Zero Bound on Interest Rate Policy.” *Journal of Money, Credit and Banking* 32 (November, Part 2): 1007–35.

______________. 2004a. “Monetary Policy in the New Neoclassical Synthesis: A Primer.” Federal Reserve Bank of Richmond *Economic Quarterly* 90 (Summer): 3–20.

______________. 2004b. “The Monetary Policy Debate since October 1979: Lessons for Theory and Practice.” Paper for “Reflections on Monetary Policy: 25 Years after October 1979.” Conference at Federal Reserve Bank of St. Louis, October 7–8.

______________. 2005. Inflation Targeting in the United States? In *The Inflation-Targeting Debate,* ed. Ben S. Bernanke and Michael Woodford. Chicago: The University Of Chicago Press: 311–37.

______________, and Robert G. King. 1997. The New Neoclassical Synthesis. In *NBER Macroeconomics Annual*, ed. Ben S. Bernanke and Julio Rotemberg. Cambridge, Mass.: MIT Press.

Hetzel, Robert L. 1998. “Arthur Burns and Inflation.” Federal Reserve Bank of Richmond *Economic Quarterly* 84 (Winter): 21–44.

______________. 2003. “Japanese Monetary Policy and Deflation.” Federal Reserve Bank of Richmond *Economic Quarterly* 89 (Summer): 21–52.

______________. 2004a. “How Do Central Banks Control Inflation?” Federal Reserve Bank of Richmond *Economic Quarterly* 90 (Summer): 47–63.

______________. 2004b. “Price Stability and Japanese Monetary Policy.” Bank of Japan *Monetary and Economic Studies* 22 (October): 1–23.

______________. 2005. “An Analytical History of the Monetary Policy of the Federal Reserve System.” Mimeo, Federal Reserve Bank of Richmond.

Hume, David. [1752] 1955. Of Money. In *David Hume Writings on Economics,* ed. Eugene Rotwein. Madison: University of Wisconsin Press: 33–46.

King, Robert G., and Alexander L. Wolman. 1999. “What Should the Monetary Authority Do When Prices Are Sticky?” In *Monetary Policy Rules*, ed. John B. Taylor. Chicago: The University of Chicago Press: 349–98.

Kydland, Finn E., and Edward C. Prescott. 1982. “Time to Build and Aggregate Fluctuations.” *Econometrica* 50 (November): 1345–70.

Lucas, Robert E., Jr. [1972] 1981. Expectations and the Neutrality of Money. In *Studies in Business-Cycle Theory*. Cambridge, Mass.: The MIT Press.

______________. [1973] 1981. Econometric Testing of the Natural Rate Hypothesis. In *Studies in Business-Cycle Theory*. Cambridge, Mass.: The MIT Press.

______________. [1980] 1981. Rules, Discretion, and the Role of the Economic Advisor. In *Studies in Business-Cycle Theory*. Cambridge, Mass.: The MIT Press.

______________. 1996. “Nobel Lecture: Monetary Neutrality.” *Journal of Political Economy* 104 (August): 661–82.

McCallum, Bennett T. 1983. “The Role of Overlapping-Generations Models in Monetary Economics.” *Carnegie-Rochester Conference Series on Public Policy* 18 (Spring): 9–44.

______________. 2002. “Recent Developments in Monetary Policy Analysis: The Roles of Theory and Evidence.” Federal Reserve Bank of Richmond *Economic Quarterly* 88 (Winter): 67–96.

______________, and Marvin Goodfriend. 1987. Demand for Money: Theoretical Studies. In *The New Palgrave: A Dictionary of Economics*, ed. John Eatwell, Murray Milgate, and Peter Newman. London: The Macmillan Press: 775–80.

Mishkin, Frederic S. 1999. “The Case for Inflation Targets: How to Make the Fed Foolproof.” *Fortune* (March 1): 52–3.

Modigliani, Franco, and Lucas Papademos. 1975. “Targets for Monetary Policy in the Coming Year.” Brookings Papers on Economic Activity 1: 141–63.

Nelson, Edward. 2003. “The Future of Monetary Aggregates in Monetary Policy Analysis.” *Journal of Monetary Economics* 50: 1029–59.

Orphanides, Athanasios, and Simon van Norden. 2004. “The Reliability of Inflation Forecasts Based on Output Gap Measures in Real Time.” Working Paper. Finance and Economics Discussion Series, Federal Reserve Board.

Porter, Richard D., and David H. Small. 1989. “Understanding the Behavior of M2 and V2.” *Federal Reserve Bulletin* 75 (April): 244–54.

Prescott, Edward C. 1986. “Theory Ahead of Business-Cycle Measurement.” *Carnegie-Rochester Conference Series on Public Policy* 25 (Autumn): 11–44.

Rotemberg, Julio J. 1987. “The New Keynesian Microfoundations.” In *NBER Macroeconomics Annual*, ed. Stanley Fischer. Cambridge, Mass.: MIT Press.

______________, and Michael Woodford. 1997. “An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy.” In *NBER Macroeconomics Annual*, ed. Ben S. Bernanke and Julio Rotemberg: 297–346.

______________. 1999. Interest Rate Rules in an Estimated Sticky Price Model. In *Monetary Policy Rules*, ed. John B. Taylor. Chicago: The University of Chicago Press: 57–119.

Rudebusch, Glen D., and Lars E. O. Svensson. 1999. Policy Rules for Inflation Targeting. In *Monetary Policy Rules*, ed. John B. Taylor. Chicago: The University of Chicago Press: 203–53.

Samuelson, Paul, and Robert Solow. [1960] 1966. Analytical Aspects of Anti-Inflation Policy. In *The Collected Scientific Papers of Paul A. Samuelson* 2 (102), ed. Joseph Stiglitz: 1336–53.

Sargent, Thomas J. [1971] 1981. A Note on the "Accelerationist" Controversy. In *Rational Expectations and Econometric Practice*, vol. 1., ed. Robert E. Lucas, Jr., and Thomas J. Sargent. Minneapolis: The University of Minnesota Press.

______________, and Neil Wallace. [1975] 1981. "Rational" Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule. In *Rational Expectations and Econometric Practice*, ed. Robert E. Lucas, Jr., and Thomas J. Sargent. Minneapolis: The University of Minnesota Press: 215–28

Wolman, Alexander L. 1997. "Zero Inflation and the Friedman Rule: A Welfare Comparison." Federal Reserve Bank of Richmond *Economic Quarterly* 83 (Fall): 1–21.

______________. 1998. "Staggered Price Setting and the Zero Bound on Nominal Interest Rates." Federal Reserve Bank of Richmond *Economic Quarterly* 84 (Fall): 1–24.

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

______________. 2001. "A Primer on Optimal Monetary Policy with Staggered Price-Setting." Federal Reserve Bank of Richmond *Economic Quarterly* 87 (Fall): 27–52.

Woodford, Michael. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton, N.J.: Princeton University Press.

Equilibrium Models of Personal Bankruptcy: A Survey

Kartik Athreya

Academic research aimed at understanding the consumer default decision has grown rapidly over the past decade. The genesis of this research is the product of three broad sets of forces. First, advances in pure theory gave economists a better understanding of the implications of allowing default for consumer welfare and credit market outcomes. Second, the relatively striking growth of unsecured consumer debt and default in the 1990s spurred the interests of applied researchers in explaining the default decision.[1] Third, and most recently, advances in computational technology have allowed economists to map the insights from pure theory into models capable of confronting observed data and yielding quantitative implications.

This article documents the evolution of recent work on personal bankruptcy. The questions addressed by this research range from the role of income uncertainty in driving financial distress to the roles of statutes allowing households to shelter wealth and of decreased moral commitment to repay debts.

The extant literature consists of a variety of economic approaches to policy analysis. For example, several recent papers employ detailed analyses of observed data. Some of these analyses—notably Gropp, Scholz, and White (1997); Elul and Subramanian (2002); and Grant (2003)—cleverly exploit the near-natural experiments provided by interstate variation in bankruptcy law. At the other end of the spectrum are the so-called "equilibrium" approaches, typified in the work of Athreya (2002, 2004, forthcoming), Chatterjee, Corbae, Nakajima, and Rios-Rull (2002); Li and Sarte (forthcoming); Livshits,

I thank Huberto Ennis, Pierre Sarte, and John Weinberg for very helpful comments. The views expressed are those of the author and do not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System. All errors are mine.

[1] Throughout this article, the words "bankruptcy" and "default" are used interchangeably.

MacGee, and Tertilt (2003); and others. These approaches have in common settings in which households optimize, and in which equilibrium conditions such as those implied by competition, market clearing, and resource feasibility are imposed. Such researchers explicitly solve household optimization problems parameterized through "calibration" or estimation, and then employ simulation to understand bankruptcy policy. What follows documents most directly the contributions of these "equilibrium" models.

The article is organized as follows. Section 1 presents the intuition captured in purely theoretical models of default. Section 2 documents some of the empirics pertaining to personal bankruptcy. Section 3 discusses a set of quantitative equilibrium models based on the theoretical foundations discussed in the first section. Each of these models analyzes the role of default at both the individual and aggregate levels in a manner that respects salient features of U.S. data. The final section concludes.

1. BASIC THEORY

The Role of Limited Insurance

Proponents of current bankruptcy law have long argued for the role of debt forgiveness in helping those hit by unexpected hard times. The "honest but unfortunate debtor" is now folkloric. Central to this view is a belief that life is characterized by a nontrivial amount of *uninsurable* risk. In particular, instead of pooling risks through explicit insurance contracts, households may do some risk-management by saving in good times and borrowing in bad times. The spectre of persistent misfortune, as it may leave households in financial straits, raises the possibility that occasionally allowing default may be useful. In other words, the default option may act like an insurance policy against very bad luck.

Formalizing the insurance role of the option to default is the subject of several recent works. The seminal references are the papers of Dubey, Geanakoplos, and Shubik (2005) and Zame (1993), who both study outcomes in stylized two-period models in which all uncertainty resolves at the terminal date. The critical assumption made in their work is that insurance markets against second-period risk are incomplete. Of course, without this imperfection, allowing default simply corrodes the ability of borrowers to commit to repaying debts, and can therefore only make matters worse. The key result of their work is that in such a world, allowing default subject to a finite penalty can improve allocations, even if it means that default occurs in equilibrium. The intuition is that default, by allowing repayment to partially suit the immediate needs of a household, can provide a "state-contingency" to debt that aids smoothing. The following elegant example from Zame (1993) makes the point more clearly.

Example 1: Default Can Be Especially Useful

The analysis in Zame (1993) contains two central insights. The first insight is that allowing default can very nearly create a fully insured income stream in situations where *completely* insuring against risk may require promising to occasionally pay very large amounts to a counterparty. The second insight is that if the underlying structure of assets is such that their payoffs are "similar" to each other, then households will necessarily be forced to promise very large payments to counterparties in some states in order to obtain insurance against other states. In such a setting, the addition of more assets with similar structure will *not* help. That is, default is uniquely useful in overcoming insurance-market incompleteness.

Consider a world with two agents with identical, risk-averse, standard, expected utility functions, and two dates, 1 and 2. On the first date, both agents are endowed with a unit of the good. On the second date, each agent draws stochastic income in the form of a single consumption good that depends on an aggregate state that takes values from a countably infinite set, $\Omega = \{1, 2, ...\}$. The endowment of Agent 1, (where the first entry is the date-1 endowment), is denoted e^1. The endowment of each agent depends on the aggregate state as follows: $e^1 = \{1; 1, 7, 1, 1, 1, ...\}$. Similarly, Agent 2 faces endowment $e^2 = \{1; 7, 1, 1, 1, 1, ...\}$. Next, we impose a probability distribution, $\mu(\omega)$, over income levels, $\omega = \{1, 2, ...\}$. Specifically, let $\mu(1) = \mu(2) = 1/4$, and $\mu(\omega) = 3^{-\omega+2}$ for $\omega > 2$. This probability distribution has the property that the probability of receiving income from a state with a relatively high index, ω, is far lower than the probability of receiving income from a state with a relatively low index.

A complete-markets Walrasian equilibrium for this economy is Pareto efficient, whereby there is perfect risk sharing between the two agents. Therefore, given the symmetry in preferences and endowments, the equilibrium allocation is one in which consumption for each agent, $c^i, i = 1, 2$, is an equal share of aggregate income. Under the assumed aggregate endowment process, equilibrium consumption is therefore given by $c^i = \{1; 4, 4, 1, 1, 1, ...\}$, for $i = 1, 2$. Notice that this means that individual-level consumption varies only with the aggregate endowment of the economy, and not with any agent-specific changes in endowments.

To achieve this allocation, Agent 1 gives up three units of date-2, state-2 consumption, in return for three units of date-2, state-1 consumption, with Agent 2 taking the other side of the transaction. Lastly, consider the following finite set of assets, A_1,A_2,...A_K, that will allow some, but not complete, insurance. Let each row in the matrix below describe the state-contingent date-2 payoffs for a given asset. The following asset structure does not permit perfect smoothing.

$$\begin{array}{rcccccccccc}
A_1 = & 1 & 2 & 0 & 0 & 0 & . & & . & . & . \\
A_2 = & 0 & 1 & 2 & 0 & 0 & 0 & & . & . & . \\
A_3 = & 0 & 0 & 1 & 2 & 0 & 0 & & . & . & . \\
. & 0 & 0 & 0 & 1 & 2 & 0 & & 0 & . & . \\
. & . & . & . & . & . & . & & . & . & . \\
. & . & . & . & . & . & . & & . & . & . \\
A_K = & . & . & . & . & . & . & \multicolumn{2}{c}{.1(k^{th}\text{ position})} & 2 & .
\end{array} \tag{1}$$

In this case, notice that to replicate the payoffs of the Pareto efficient outcome in the first four states, for example, Agent 1 needs to hold the following portfolio:

$$F_4 = 3A_1 - 9A_2 + 18A_3 - 36A_4.$$

While this allocation replicates the Pareto optimum, by construction, for states $1-4$, the liability that Agent 1 incurs in state 5 is -72 units of consumption. This liability far exceeds Agent 1's endowment in state 4 (which is 1). If default were disallowed, then the given asset structure could not replicate the optimal allocation for even the first four states. Moreover, the essence of Zame's example is that the addition of more assets with the payoff structure here beyond $A_1 ... A_K$ will not help, so long as we continue to require that all liabilities be satisfied with certainty. As seen, with securities $A_1 ... A_{2N}$, to achieve the optimal risk sharing for states $1 - 2N$, Agent 1 would owe -9(2^{2N-1}) in state $2N + 1$. Given the endowment structure, this is again infeasible. However, if debt forgiveness is allowed, with a finite penalty, λ—that is, assumed proportional to the liability—matters change. For default in state $2N$, the expected penalty is $9\lambda(\frac{2}{3})^{2N-1}$. Therefore, the possibility exists for λ small enough that Agent 1 will be willing to hold portfolio F_{2N}, while Agent 2 finds this acceptable. The key, to repeat, is that default can void the need for households to hold very large liabilities, and can also facilitate trades that additional assets, however many, may not be able to achieve.

The Role of Limited Commitment

The opponents of the above arguments argue that any benefits that potentially arise from default are confronted by even larger costs. Perhaps most commonly, opponents of default and bankruptcy have argued that debt forgiveness and other forms of limited liability may simply encourage profligacy, sloth, and impose costs on other, more judicious borrowers.[2] Moreover, lenders and insurers themselves may be wary of entering into contracts with households

[2] Arguments against bankruptcy tend to center around the spill-overs that may accompany default. For example, the argument that easy default reduces thrift is only germane if it changes the opportunities that others may have. Similarly, if easy bankruptcy encourages shirking at the

endowed with the right to "walk away." In other words, an important possibility is that it is misleading to take incompleteness as a given when studying bankruptcy, precisely because *incompleteness may be caused by the option to default* in the first place. The crux of this argument is twofold. First, the willingness of households to commit to repayment is driven by, among other things, the extent to which they can self-insure. If bankruptcy offers good self-insurance, then participation and the demand for formal insurance may be diminished. That is, private credit with default may be crowding out formal insurance, leading observers to wrongly conclude that markets for risk remain highly incomplete. This strand of the literature, typified by the theoretical work of Kocherlakota (1996), begins by allowing a full set of insurance contracts to be traded, and then studies the extent to which limited commitment for repayment "shrinks" the feasible set of contracts. In this manner, these models produce incomplete insurance as an *outcome*, as opposed to merely asserting it to begin with. In these settings, because insurance markets are rich enough to allow full risk sharing under unlimited liability, the welfare consequences of introducing consumer bankruptcy are unambiguously negative. The following simplified example illustrates the manner in which the possibility of default limits insurance provision.

Example 2: Default Can Cause Incomplete Insurance

Consider a risk-averse consumer who faces income risks. Assume that the consumer can contract with a perfectly diversified insurance company who *is* committed to honoring all contracts. Denote a finite set of income realizations by $y \in \{y_1, y_2, ...y_N\}$, where $y_i < y_j$ if $i < j$. Denote the relative likelihoods of various income realizations by a probability distribution π_1, π_2,...π_N, whereby $\sum_{i=1}^{N} \pi_i = 1$. Let household utility be given by $u(C(y_i))$, where $C(y_i)$ denotes consumption in state-i. Let mean income be normalized to one, and let $P(y_i)$ denote the transfers made by the consumer to the insurance company. Negative values are interpreted as transfers to the consumer, while positive values represent payments by the consumer.

In this setting, the problem of the household is the following:

$$\max \sum \pi_i u(C(y_i)),$$

s.t.

$$C(y_i) = y_i - P(y_i).$$

workplace, or reduces search effort when unemployed, this too is relevant for social well-being only to the extent that such behavior places a burden on an unwitting employer or unemployment insurance scheme. The ability of bankruptcy to impose such costs depends fundamentally on what aspects of individual behavior may be observed—and, hence, priced—in competitive markets. For example, if shirking on the job and search effort among the unemployed were costly to observe, the distortion to behavior created by a default option might be socially quite harmful. I will return later to the role of observability in altering the desirability of allowing bankruptcy protection.

The solution to this problem is perfect insurance. Namely, the consumer will be left, net of insurance premiums, with a perfectly smooth profile of consumption equal to the mean income of unity, regardless of the income realization. To achieve this profile, the consumer will pay a premium equal to the excess of current income over mean income: $y_i - 1$, whenever income is larger than the mean, while receiving the difference between realized income and mean income whenever the former is smaller.

The limited commitment is now introduced as follows. Assume that the consumer cannot credibly commit to repay any more than a fraction, $\psi < 1$. Intuitively, it may help to think of a consumer who can leave town if asked to pay more, and, further, that the consumer has no way of *credibly* promising to the insurance company that he will remain in the contract. To facilitate comparison with the full-commitment problem, we see that the household still solves:

$$\max \sum \pi_i u(C(y_i)),$$

s.t.

$$C(y_i) = y_i - P(y_i).$$

However, an insurance contract must now satisfy:

$$P(y_i) \leq \psi y_i,$$

which represents limited commitment or liability. While the solution is slightly cumbersome and, therefore, not presented here, the intuition for the solution of this problem is straightforward.[3] If the insurance company wishes to break even, it must not contract with the consumer in a way that leaves it vulnerable to default. In other words, it must not rely on large payments from the consumer in good states to offset payments it makes to consumer in bad states. Instead, it must *limit* payments in bad states in a way that allows it to break even, given that it will not be able to collect in good states. In other words, limited commitment can lead to incomplete insurance.

Despite the possibility that limited commitment may create market incompleteness, a caveat is in order. One must be careful to distinguish the repudiation of a true "insurance" contract, whereby payments are negatively correlated with one's income, from repudiation of a "debt" contract, whereby payments are uncorrelated (up to bankruptcy) with one's income. Notably, default in the former occurs when income is *high,* while default in the latter occurs when income is low. Moreover, even if it is limited commitment that is responsible for incomplete insurance, it may still be useful to have default as an option. In particular, if there was incompleteness of insurance contracts,

[3] The interested reader is referred to Obstfeld and Rogoff (1996, Chapter 6) for the detailed solution to this problem and to Ljungqvist and Sargent (2000, Chapter 15) for the more general dynamic limited commitment problem modeled on Kocherlakota (1996).

regardless of whether incompleteness arose from limited commitment, the incentive for households to trade contracts to share risk (e.g., through assets such as those specified in Zame's example) would remain.

Default and Moral Hazard

In the previous example, information was assumed to be perfect, and the incompleteness of insurance emerged solely from the inability of the insured to commit to the full-insurance contract. If enough information is available to limit opportunistic behavior by insured households, bankruptcy will be essentially unnecessary. Moreover, in such a setting, any distortions to allocations caused by bankruptcy will be limited to the level of borrowing that may take place, since the interest rate on loans will need to reflect default risk. If creditors have few ways of punishing defaulters, as appears to be the empirically relevant case, one might expect default risk to be high in such a world. In particular, exclusion from borrowing, the primary long-run consequence of bankruptcy, may not be an effective deterrent to default when insurance markets are complete. However, purely intertemporal smoothing, such as that undertaken by the young to finance education, will still be greatly hindered if borrowing is very expensive. Such a potentially important distortion may persist even with relatively rich insurance markets.

By contrast, if information about the actions taken by insured households is not easily available, then bankruptcy may impose costs that are rather large. In the context of unemployment insurance schemes for example, bankruptcy may greatly exacerbate the incentives to shirk and even take on additional risks. That is, by enhancing the ability to self-insure, the provision of explicit insurance may become undesirable. This is an instance of the more general problem of "moral hazard," whereby an insured party's incentives to take risks in a manner unobservable to the insurer are enhanced. In recent work, Athreya and Simpson (forthcoming) argue that as a quantitative matter, the ability of the U.S. public unemployment insurance system to improve welfare is seriously limited by the availability of bankruptcy.

In sum, the principal insights of the theoretical work presented above are as follows:

- *Default may enhance the functioning of asset markets in helping consumers hedge risks.*

- *Limited insurance, though it may justify allowing the option to default, may itself arise from limited commitment.*

- *The availability of default can be expected to increase moral hazard.*

2. SOME FACTS

Beginning in the late 1980s and early 1990s, events in unsecured credit markets attracted the attention of economists. It was a period characterized by three features: increased credit availability, increased indebtedness, and rapidly increasing personal bankruptcy rates. Striking facts such as these begged explanation, and prompted several analyses. With respect to the trends mentioned above, the interested reader is referred to Sullivan, Warren, and Westbrook (1989, 2000). This article, however, will focus less on accounting for these facts and more on documenting aspects of bankruptcy that have remained more stable over time. In particular, the scrutiny triggered by the growth of credit and default-related variables led to the discovery of a variety of other facts that seem to be relatively time-invariant. It is these more "long-run" phenomena that are the subject of several recent papers discussed below.

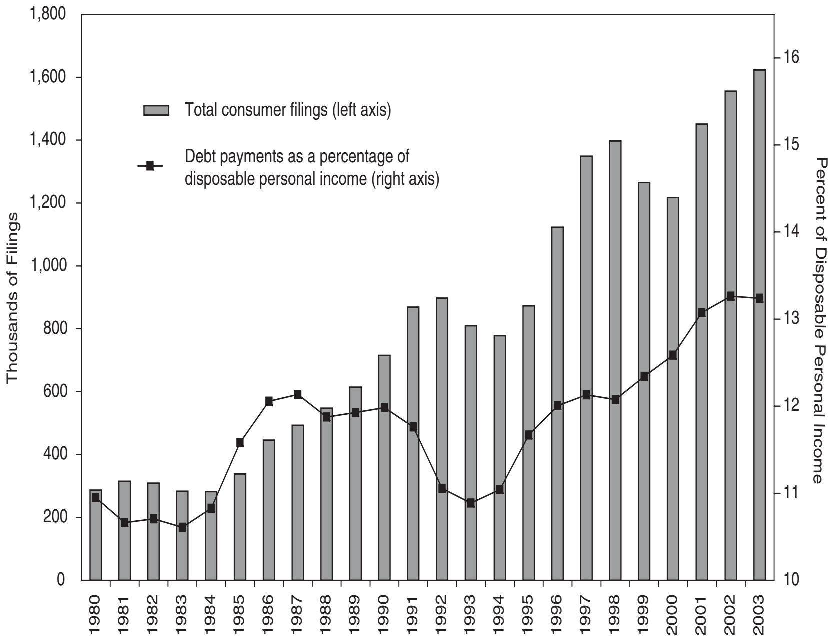
The Facts: Bankrupts "R" Us, but Not Quite

Are bankruptcy filers educated or uneducated? Are they rich or poor, young or old, sick or healthy, employed or unemployed, entrepreneurs or workers? The extant body of work has taught us much about "who" bankruptcy filers are. The starting point in research on the demographics of bankruptcy filers is the seminal work of Sullivan et al. (1989, 2000). The 1989 study surveyed bankruptcy filers in the 1980s, while the second captured data through 1997.

With respect to educational attainment, the distributions of attainment are strikingly similar, (Sullivan et al. 2000, 53) with bankruptcy filers even more likely than the general population to report having attained "some college." However, this is potentially misleading in that bankruptcy filers are somewhat less likely than the rest of the population to hold either a college or advanced degree.

More interesting differences appear when comparing the earnings of filers and non-filers for a given level of educational attainment. In Sullivan et al. (2000), the earnings of college-educated non-filers to filers is 1.4, and is most extreme for those with advanced degrees, where the ratio is roughly 2. This fact suggests that bankruptcy filers are disproportionately the recipients of degrees which offer lower returns. Foreseeable aspects of human capital acquisition may be the cause of this situation as well as the possibility that bankruptcy filers are often those for whom the uncertain nature of post-college earnings have been resolved in favor of poor earnings. Lastly, in the year that a filing occurs, incomes among filers are substantially lower than for non-filers. For example, a sample of five U.S. states shows that the median income for filers is often only half that of non-filers, while the mean is even more skewed in favor of non-filers. Moreover, recent work of Sullivan et al. (2000) and Bermant and Flynn (2002) suggests that the unemployment rate among filers is between three and four times as high as is for the rest of the population,

Figure 1 Influence of Total Consumer Debt on Bankruptcy Filings Trends by Year, 1980–2003



giving fillip to the narrative that places shocks to earnings front and center in explaining personal bankruptcy.

Net indebtedness is much higher for the population of filers than for the rest. The only reason that indebtedness is not an obvious feature of the data is that the median asset exemption as of 1998, measured by Grant (2003), was $44,000. In other words, households could hold large stocks of wealth while holding unsecured debts that were dischargeable in bankruptcy. Nonetheless, few households in bankruptcy hold large stocks of wealth. Median net worth among filers was -$10, 500 in 1991, while it was $36,000 for the U.S. population as a whole. However, the income levels reported above suggest that it may simply be the mismatch between income and debts, rather than the size of debts, that matters for bankruptcy. Figure 1 is suggestive of the role of debt relative to income in accounting for recent trends in bankruptcy.

With respect to age, bankruptcy filers are also similar in age to the general population. Sullivan et al. (2000) find in their sample that, as of 1997, the median age of filers was only slightly lower than the national median of 41.7 years. However, the median hides the feature that roughly 85 percent of filers are younger than 55. A final dimension along which bankruptcy filers may be distinguished is by their prior employment experience. Specifically, entrepreneurship is featured prominently in the histories of bankruptcy filers. Sullivan et al. (1989) document that fully 20 percent of filers in their data from the 1980s report themselves to be self-employed in a business venture, even though entrepreneurs account for only slightly more than 10 percent of the U.S. population. In sum:

Bankruptcy filers are similar to the overall U.S. population along the dimensions of education and median age, but differ from the overall U.S. population in that are they more likely to be young, low net-wealth, sick, unemployed, and self-employed.

3. QUANTITATIVE THEORY: RUNNING A HORSE RACE

More recently, the theoretical contributions of Dubey et al. (2005), Zame (1993), Kocherlakota (1996), and others have been critical in motivating dynamic general equilibrium models aimed at evaluating the consequences of bankruptcy. The distinguishing features of these models are much greater attention to institutional richness pertaining to personal bankruptcy statutes and a reliance on numerical solutions. Notably, the households in these settings can be distinguished clearly by age, debt portfolios, insurance policies, and other ways observable in the data. The payoff to the introduction of these complications is to allow for *quantitative* evaluations of the relative strengths of the forces of incomplete insurance and limited commitment.

To provide intuition as to how the particular quantitative aspects of incomplete insurance operate, consider the following: As a quantitative matter, the relevance of Dubey et al.'s (2005) justification depends on the precise difficulties imposed by incomplete insurance markets that, in turn, depend crucially on the nature of income risk faced by households. Labor income is important as it accounts for most household income and is typically not directly insurable (perhaps for the obvious complications created by moral hazard). For example, consider an environment in which households that face risk to income may borrow and save, but have no option to default. That is, all debts have to be honored, regardless of the circumstances facing the borrower. Suppose also that the household was susceptible to both short-term and very long-lasting shocks. Relative to a short-term shock, receiving a long-term shock can much more seriously alter the present value of future income. If such a shock were negative, the present value of future income could fall substantially, making the acquisition of debts to smooth consumption in the face of short-term shocks potentially very expensive.

A key implication of standard models of consumption and savings is that households borrow (or "dissave") progressively less as shocks grow more persistent. In other words, the willingness of households to effectively self-insure temporary income disturbances might reasonably depend on their likelihood of receiving more persistent shocks. This is one sense in which the possibility of default can "grease the wheels" of credit markets. This line of reasoning is addressed more fully in Athreya and Simpson (forthcoming).

The relative strength of the effects of the forces listed above are ultimately key to deciding on the possibility of a beneficial role for bankruptcy. While there is a good deal of empirical work describing the salient facts, questions of welfare are not addressed in that literature. The interested reader is referred to the thorough review of empirical work contained in the Congressional Budget Office (2000). I focus instead on a handful of representative quantitative models, all of which contain forward-looking households that face uninsurable risks. An advantage of this approach is that welfare can be easily evaluated for equilibriums allocations. The first group of environments contain a single asset, attached to which is a bankruptcy option. These models are essentially direct extensions of the work of Aiyagari (1994) and Huggett (1993). However, the payoffs from the single asset do not vary for any other reason, which precludes analysis of aggregate risk and also constrains insurance possibilities in a particularly strong manner. The second class of models allows for more than one asset, as well as for richer insurance possibilities beyond noncontingent unsecured consumer debt. For convenience, Table 1 presents a stylized taxonomy of the models discussed below.

Single-Asset Models

Chronologically, the first quantitative evaluation of bankruptcy appears to be Zha (2001). The key features of this environment are (1) a large number of risk-averse households and (2) the presence of uninsurable, but purely idiosyncratic, risk. That is, aggregate activity in this economy does not fluctuate through time, making it incapable of addressing questions related to business cycles. Households are free in this setting to save by accumulating capital, which in turn is used to produce output, or by borrowing to invest in (idiosyncratically) risky capital equipment to produce output. Borrowers operate investment projects that allow this uncertain output to be seen only at a cost.

An important assumption made in Zha (2001) is that contracts are static, a feature used later in several other papers (e.g., Athreya [2002, 2004, forthcoming]; Livshits et al. [2004]; Chatterjee et al. [2002]; and Li and Sarte [2002]). It is known from Townsend (1979) that in a setting where a lender faces a cost of verifying the "state" (income, in this case), the optimal static contract is one resembling "debt"—that is, a contract where the borrower announces output (truthfully) and repays a constant amount unless output is relatively low. In such cases, the household announces (truthfully) that output is low, and the lender seizes a fraction of the output. The amount of wealth that may be given

Table 1 Comparing Some Equilibrium Models

	Athreya (2002)	Athreya (2004, forthcoming)	Livshits et al. (2004)	Li/Sarte (2004)	Chatterjee et al. (2002)	Mateos-Plannas/Seccia (2004)
Credit Limits/ Observability	Exogen./Only Unconditional Moments	Endogen./ Total Debt	Endogen./ Total Debt, Current Prod., Age	Exogen./Only Uncondtional Moments	Endogen./ Total Debt, Current Prod.	Endogen./Only Unconditional Moments
Expense/ Preference Shocks?	No	No	Yes	No	Yes	No
Flexible Labor Supply/ Production?	No	No	Yes	Yes	No	No
Can the "No-Bankruptcy" Case Be Studied?	Yes	No/Yes	No	Yes	No	Yes
Finite Horizon or Infinite Horizon?	Infinite	Infinite	Finite	Infinite	Infinite	Infinite
Allows Chapter Choice?	No	No	No	Yes	No	No
Single or Multiple Asset?	Single	Single/ Multiple	Single	Multiple	Single	Single
Partial Equilibrium (PE) or General Equilibrium (GE)?	GE	PE	PE	GE	PE	GE

up to satisfy unsecured creditors is restricted in practice by "exemptions" on bankruptcy. Zha (2001) finds that increasing the value of exemptions does have a positive effect on welfare. Zha (2001) does not, however, compare the welfare consequences of repealing all default to a setting with default.

Bankruptcy reform efforts as of the late 1990s centered around two aspects of the law. First, existing bankruptcy law did not require "means-testing" for would-be filers. Therefore, potentially high-income filers would be able to discharge substantial debt without giving up income. Note carefully the distinction between income and wealth. Exemptions are rules that explicitly cover the levels of wealth that debtors may retain in bankruptcy. Means-tests did not directly apply to these provisions. Athreya (2002) addressed the issue of means-testing in a simpler environment than that studied by Zha (2001). The key features of the environment, as in Zha (2001), were a large number of infinitely lived households facing income risk. Unlike Zha (2001), creditors were assumed to issue credit cards with fixed "lines" of credit, where the interest rate on loans did not move with the total debt level. The advantage of this assumption is that contracts were simple fixed-rate instruments that bore a closer resemblance to the predominant form of unsecured debt held by U.S. households. The disadvantage is that credit conditions are restricted to appear only through prices, whereby bankruptcy law would have no effects on credit limits.

Athreya (2002) finds that the means-testing provisions embodied in the Bankruptcy Reform Act of 1999 would not greatly alter allocations. The intuition is that, given the costs of bankruptcy implied by the data, most households would not file unless already poor. On the other hand, in Athreya (2002), the elimination of all bankruptcy was found to be quite beneficial. This is important because, as mentioned earlier, the main costs of bankruptcy in Athreya (2002) come from (1) interest rate "externalities" arising from the pricing of loans to cover average, as opposed to personalized, repayment rates and from (2) the assumption that all costs of bankruptcy were deadweight in nature. In other words, easy bankruptcy law generated a shift of households toward the use of bankruptcy that necessitated the more frequent use of deadweight penalties. On balance, the frequency of filing under lax penalties for bankruptcy did not help households smooth consumption in a manner that offset society's cost of imposing court costs and possibly "stigma" on filers.

The estimate in Athreya (2002) is, however, prone to producing a downwardly biased estimate of the benefits of eliminating default, at least for the income process employed.[4] This bias arises because with only marginally more observability, whereby creditors could see total household debt, one would have expected a large reduction in the cost of credit. Therefore, if the initial credit limits assumed in Athreya (2002) were lax enough that the expansion in credit did not exceed the limit initially assumed, then the estimate may be reasonable. However, if the initially assumed debt level was narrower than the endogenous debt level that would emerge under partial observability of debt, then the estimate of the benefits of strict bankruptcy are too low.

Recognizing that the willingness of creditors to lend will be related to the characteristics that they may observe about a borrower, Chatterjee et al. (2002), Livshits et al. (2004), and Mateos-Planas and Seccia (2004) each allow for more observability than Athreya (2002). Chatterjee et al. (2002) use an infinite-horizon setting where current period income and current wealth are both observable, in principle. In addition to allowing for endogenous limits on debt, Chatterjee et al. (2002) also contribute by establishing theoretical properties associated with recursive representations of household optimization problems, specifically for the case of i.i.d. income shocks. Livshits et al. (2004), on the other hand, use a life-cycle model and augment the set of debtor observables to include age. In contrast to both preceding papers, Mateos-Planas and Seccia (2004) assume a lack of observability beyond population averages. They appeal to an institutional structure whereby banks finance lending by issuing securities backed by repayments on the unsecured

[4] We revisit this issue when discussing the alternative earning/expense process used by Livshits et al. (2004).

loans they make. The "pools" backing these securities are large and aggregate a variety of types of households. This route was originally taken in Dubey et al. (2005) in order to conform to the predominant form of financing used by credit card and mortgage lenders. Mateos-Planas and Seccia (2004) then go beyond the exogenous limits to borrowing used in Athreya (2002) and derive borrowing limits consistent with zero profits under this restriction on observability.

The policy experiments pursued in Chatterjee et al. (2002) study the desirability of (1) means-testing and (2) a reduction in the length of time for which credit reporting agencies may retain the record of a bankruptcy filing. They find that nearly all households would prefer to restrict the option of bankruptcy to those households with above-median earnings, and few at all are in support of a reduction in the legal limit on the length of time that a past bankruptcy filing may be recorded. Similarly, with respect to whether bankruptcy should be allowed or not, Athreya (2002) finds that prohibiting bankruptcy is preferable to permitting it as currently practiced. The intuition for these findings is that the increased costs of credit apply to all borrowers (although not with equal force) and is an externality that offsets the better consumption insurance allowed by bankruptcy in isolation. Athreya (2002) also finds that means-testing is not very important, while the availability of bankruptcy is. Notably, at present there are few restrictions on the current earnings of a household at the time of a Chapter 7 bankruptcy filing. Nonetheless, the income of filers is rarely higher than the median.

Athreya (2002) abstracts from certain types of catastrophic risks, such as severe health shocks, legal costs of divorce, and the inability to collect mandated child support. The shock processes used in the preceding models affect only income. However, to the extent that such severe shocks are important, as Sullivan et al. (2000) argue is true for approximately 20 percent of all filings, income shocks underrepresent the risk faced by households. Livshits et al. (2004) therefore propose a more extreme feature—that of shocks to "expenses," or net worth. Households are assumed to be subject to occasional, but large, reductions in their asset positions. These shocks are meant to capture large unanticipated expenditures on inelastically demanded types of goods, such as hospital care or child care. The advantage created by this is to allow for real risks that appear important in at least a portion of personal bankruptcy filings (for example, 19 percent of filings involved large medical expenses, according to Sullivan et al. [2000]). However, the presence of expense shocks that may make current resources deeply negative under the assumption of standard constant-relative-risk-aversion preferences implies that without the possibility of default, no one would borrow. This extreme outcome arises because the full enforcement of debts would mean forcing the household to zero consumption, i.e., literally forcing it to starve in the face of large negative shocks. Knowing this, the household would never *voluntarily* assume debts.

The disadvantage created by this condition is that the comparison of outcomes with and without a bankruptcy option is not easy to evaluate. On the other hand, the results of Livshits et al. (2003) are a useful, if stark, benchmark for how the availability of default facilitates risk-sharing through asset-markets. Namely, since prohibiting default in their environment would mean not borrowing at all, a portion of the welfare gains from bankruptcy in their model comes from the complementary role that default plays in enhancing the value of asset markets.

Multiple-Asset Models

As mentioned earlier, all of the preceding models feature a single asset that when held in negative quantities, implied borrowing. The absence of any distinction between assets means that borrowing and negative net worth are the same thing. This restriction does not match well with the complicated portfolios households have in the data. For example, Sullivan et al. (2000) document that while many households in bankruptcy do have substantial unsecured debt, many also do not have negative net worth. In large part, this restriction of trade to a single asset was driven by tractability. Some recent work has relaxed this restriction. The payoffs to accommodating this richness is that it (1) allows for a closer mapping from model outcomes to observable data, and (2) allows us to correctly measure the ability to smooth consumption and (3) allows us to analyze policies aimed at allowing households to retain wealth while discharging debts.

Tractability

With respect to tractability, the chief complication of solving a multiple asset problem is that of storing all the information the household needs to solve its optimization problem. Specifically, dynamic programming techniques are typically the tools to solve household optimization problems in environments complicated by uncertainty and "discrete" choices (e.g., to file or not to file). Dynamic programming becomes radically more cumbersome as the amount of information required of the household increases. In the context of bankruptcy exemptions, for example, a household would like to know its feasible options as it enters a period. However, these options may depend on the precise composition of household debts and assets. Moreover, even if one is creative and can define a single summary statistic that allows households entering a period to solve their optimization problem, it may be at the expense of complicating the "within-period" problem of the household. For example, in Athreya (forthcoming), households may hold secured and unsecured debt. Nonetheless, "total wealth" serves as the single variable that households must be aware of within any period.

Analyzing Exemptions: The Payoff to Allowing Multiple Assets

Despite the computational burdens involved, multiple-asset models are valuable, as they allow for the study of households that may be saving and borrowing simultaneously. In turn, such models allow researchers to meaningfully study important policy questions pertaining to bankruptcy, most notably, the rules defining exemptions. The latter govern the extent to which wealth may be held while debts are discharged. Exemptions have attracted a great deal of attention recently, both academically and in the public discussion on bankruptcy. Exemptions are, most generally, allowances for certain types of wealth that bankruptcy filers may retain after bankruptcy. Any wealth in excess of exemptions must be surrendered to satisfy unsecured lenders.

The idea that substantial wealth may be protected from seizure by creditors has long been controversial (see, e.g., Moss [forthcoming]). In support of exemptions is the following intuition: Because they partially govern how much "state-contingency" is truly embedded in an unsecured loan, it may be useful for a household to be able to hold at least one asset with payoffs that it can manipulate to serve its needs. For example, even if a loan is not explicitly collateralized, it may be implicitly so, simply because a household in bankruptcy would be required to transfer any nonexempt wealth to unsecured creditors. Thus, an unsecured loan may be effectively collateralized, even if wealth is not explicitly pledged by the debtor. By contrast, a large exemption will allow most unsecured borrowing to be cleanly discharged in bankruptcy if the household finds itself in difficult circumstances. Proponents also argue that, after bankruptcy, it is important to allow a household to run its affairs without becoming destitute and, potentially, a recipient of publicly funded transfers (see, e.g., Baird [2001]). Opponents, on the other hand, argue that those who are wealthy along some dimensions (though not along net-worth perhaps) should not be excused from debt obligations. A final concern with more ambiguous welfare implications is that to the extent that they are aware of a household's assets when making unsecured loans, creditors will price this risk. If observability is low, all unsecured borrowers will face higher borrowing costs, and if not, wealthy households alone will be able to obtain unsecured credit. We will return to the role of observability later in the article.

In a model that allows for simultaneous holdings of both debt and equity, Athreya (forthcoming) finds that, conditional on allowing bankruptcy, high exemptions are actually useful, even if they make unsecured debt more expensive. A ramification of high exemptions is that the cost of unsecured debt will be higher under high exemptions. This is precisely what data analyzed in Gropp et al. (1997) suggest. Their work is an important study on exemptions that utilizes the natural experiment provided by interstate variations within the United States. They document strong evidence supporting the view that exemptions make unsecured credit more expensive. More recently, Grant (2003) also exploits interstate variation in exemptions and finds significant support for the risk-sharing role of exemptions. His estimates indicate that exemptions noticeably reduce growth in the cross-sectional variation of income. The latter can be interpreted as an improvement in "market completeness," or insurance possibilities (see Deaton and Paxson [1994] for details).

Because bankruptcy exemptions vary across states, Elul and Subramanian (2002) document the extent to which households move across state lines to avail of more generous bankruptcy exemptions. Given that bankruptcy filers typically have low incomes at the time of filing, interstate moves may be a useful option for a household, given the fall in the market value of its time. The authors find that "forum shopping" (the explicit decision to search for a friendly set of laws) accounts for roughly 1 percent of all interstate moves to a state with a higher exemption.

Two additional recent studies deserve discussion. First, Li and Sarte (forthcoming) employ a setting in which two simplifications used in all prior work are relaxed. Their first innovation is to model the choice of bankruptcy "chapter," and the second is to accommodate production decisions. With respect to the former, in the discussion so far, I have implicitly combined all forms of personal bankruptcy. However, there are typically two forms of bankruptcy available to a household debtor. These are Chapters 7 and 13 of the U.S. Bankruptcy Code. The former is most familiar, and simply removes all unsecured debt in exchange for the surrender of all wealth above the exemption. Chapter 13, by contrast, is (at least in principle) less extreme. In particular, Chapter 13 is a form of debt rescheduling whereby a debtor agrees to repay a portion of his unsecured debts over time. This form of bankruptcy is particularly useful for households that hold wealth substantially in excess of the exemption allowed to them. However, the repayment plan poses two challenges. First, repayment over time may act like a tax on effort, and lead households to change effort. To capture the consequences of such reductions in effort, it is useful to study settings which allow for production of output. Second, current law allows households to convert a Chapter 13 filing into a Chapter 7 at any time. The latter clearly limits the extent to which repayment can be extracted from debtors. Li and Sarte (forthcoming) find notably that allowing production and capital accumulation overturns the stark results of Athreya (2002) in that eliminating bankruptcy lowers welfare. To elaborate, easy bankruptcy lowers precautionary savings and thereby lowers output, while eliminating Chapter 7 bankruptcy altogether hinders risk sharing to an inefficient degree, as Chapter 13 does not provide the same allowance for contingent repayment.

As Li and Sarte (forthcoming) allow for production, the accumulation of capital augments production and also allows for an additional asset with which households may smooth consumption. At the household level, wealth is still described by financial claims, as the capital in their model is useful only for producing the consumption good, and does not itself generate a flow of services. However, a key aspect of many exemptions applicable to bankruptcy is that they often apply only to special classes of nonfinancial assets, which typically provide a flow of services to households. The largest and most famous exemptions are those applying to home equity (the Homestead Exemption), to equity in cars, and to "tools of trade." Pavan (2003) moves research forward by explicitly modeling the services that durable goods provide their owners. This allows for more precise welfare analysis. Unlike Athreya (2002, 2004), Pavan uses a life cycle model, and estimates its parameters. Pavan's work is noteworthy for its emphasis on the use of formal statistical inference in guiding the selection of parameters. By contrast, while Athreya (2002, 2004, forthcoming), Li and Sarte (forthcoming), and others do assign values to parameters, the procedures used are less informed by formal statistical practice, and rely more on the informal matching of key features of the data. Unlike Li and Sarte (forthcoming) and Athreya (2002), however, Pavan uses a partial equilibrium model whereby the costs of funds are held fixed. Given that Li and Sarte (forthcoming) find that ignoring general equilibrium is not innocuous for studying exemptions, this distinction is worth keeping in mind.

The Role of Observability

The "insurance" value of bankruptcy depends critically on the extent to which debt prices do *not* vary with default risk. To see why, consider a world in which competing creditors were able to view a common set of factors associated with a debtor's default probability. Competition requires that cross-subsidization of some borrowers by other borrowers must not occur along any dimension that is commonly observed by creditors. This is simply because if any two borrowers have observably different characteristics, the relatively less risky borrower would be offered a cheaper rate.

As discussed earlier, assumptions on observability matter for predicting the response of lenders in the face of changes in bankruptcy policy. Therefore, these assumptions must also matter from a welfare perspective. The intuition is as follows. An insurance contract works by allowing the buyer to diversify risk and thereby transfer purchasing power from contingencies where additional consumption is valued less, to those contingencies in which it is valued more. Accurate risk-based pricing and competition will make the default option more expensive, as pricing will more closely reflect marginal default risk as opposed to the average risk that would be accounted for in the absence of such observability. However, the benefit of strong observability is that explicit insurance contracts can play an important role, voiding the need for households to rely on the relatively clumsy implicit insurance of credit with a default option.

The work of Edelberg (2003) is the first to document the changes in the pricing of credit to reflect risk at the household level. Edelberg (2003) uses data primarily from the Survey of Consumer Finances (SCF), which contains detailed information on household balance sheets, especially the level and terms of borrowing. She finds that the period beginning in the mid-1990s witnessed a sharp increase in the cross-sectional variance of interest rates charged to consumers. This work casts some doubt on the results arising in models such as Athreya (2002), where a single interest rate on unsecured credit applied to all households. According to Moss and Johnson (1999), the latter approach may have been a reasonable assumption for the 1980s and early 1990s, but advances in recordkeeping and other intermediation technologies have dropped the costs of differentiation of borrowers. The issue of intermediation costs will play an important role, as we will see in the next section. Before proceeding, however, it is useful to summarize the following provisional conclusions reached by the works cited above:

- *Under income processes that allow for large shocks to net worth, bankruptcy can play a role in improving welfare, but not without them.*
- *As a quantitative matter, moral hazard needs to be taken seriously in evaluating bankruptcy provisions.*
- *Exemptions tend to distribute credit away from the asset-poor to the asset-rich, but can improve welfare.*
- *Observability is an important determinant of how the supply side of credit markets and household welfare respond to bankruptcy policy.*

A Quantitative Equilibrium Approach for Explaining Recent Trends

Despite the now relatively large quantitative equilibrium literature on bankruptcy, the project of accounting for the time-series of the 1990s has thus far been almost exclusively tackled via purely empirical approaches. A main reason is that only recently has quantitative theorizing on bankruptcy become tractable, but even within these models, long-run "stationary" states have proved far easier to characterize than ongoing aggregate dynamics such as those observed in the 1990s. In particular, the technical problems facing quantitative equilibrium approaches in dealing with the 1990s are twofold. First, when aggregates move over time, so will prices such as the interest rate on savings or loans. Unfortunately, incompleteness of insurance is a precondition for evaluating the trade-offs associated with default. These models, in turn, contain households whose fortunes diverge with time, and as a consequence, feature wealth holdings that diverge through time. Precisely because households become heterogenous in their wealth holdings, their response to individual level or aggregate uncertainty also becomes heterogenous. In turn, future prices, which depend on aggregate wealth accumulation, are determined by the entire distribution of wealth and make problems very hard to analyze numerically. Nonetheless, Athreya (2004), described below, makes a first, and admittedly simple, attempt to account for the 1990s.

The Roles of Stigma and Technology

Detecting stigma matters for the welfare analysis of bankruptcy statutes. Let stigma be defined to mean all costs of social disapproval associated with filing for bankruptcy. Stigma matters because it is a penalty suffered by filers *after* a filing has taken place, and most importantly, one that hurts the filer, but does not directly help anyone else. If bankruptcy is an activity that society seeks to limit, then society must be aware of the possibility that if a penalty is too weak at the individual level to deter bankruptcy, imposing it very often may make it undesirable. Conversely, severe censure of filers by society may be bad for at least two reasons. First, if the incidence of bankruptcy does not fall substantially, society will find itself imposing a large "after-the-fact" punishment far too frequently. Second, if bankruptcy has a potential role in risk sharing, severe social sanctions may stunt consumption smoothing. The optimal social stigma strikes a balance between these two concerns. The preceding argument is also normative in the sense that it may tell us how strictly we wish to deal with bankruptcy filers along dimensions that merely punish them without helping others.[5]

With respect to the time path of stigma, the stunning rise in per-capita filing rates in bankruptcy, from 0.3 percent in 1980 to 1.6 percent at present, has captured the attention of many researchers. Gross and Souleles (2002) is perhaps the best known of these studies. The essence of their exercise is to ask if debtors who look similar along a multitude of financial dimensions have differentially large likelihoods of filing for bankruptcy recently relative to the past. They find the answer to be "yes," and conclude that falling stigma is a plausible story for recent data. Similarly, Fay, Hurst, and White (2002) argue that, even after controlling for state-level fixed effects, a rise in the bankruptcy rate in a given state predicts a further increase in the following year. Interpretation of empirical regularity is tricky, however. In particular, to the extent that households learn the bankruptcy process from each other and decide that it is easier than they believed, currently high filing rates may lead to even higher filing rates. From the viewpoint of predicting bankruptcy rates, whether one interprets the data as evidence of either falling stigma or learning matters little.[6] However, it may matter much more from a welfare perspective, for the reasons discussed above.

[5] The obvious analogy here is to crime and punishment, with its focus on rehabilitation against deterrence.

[6] Indeed, Gross and Souleles (2002) do not restrict their interpretations to falling stigma, but allow for the possibility that information flow has improved.

Athreya (2004) addresses the issue of stigma via an alternative route, by studying a quantitative, dynamic equilibrium model of borrowing. The article conducts two experiments. In the first case, stigma is initially "calibrated," i.e., set to a level that allows the model to match debt and bankruptcy filing rate data as of 1991. The value of this stigma is then lowered to zero. In the second experiment, stigma is fixed at the level consistent with data for 1991. The transactions costs associated with unsecured consumer lending is then sharply lowered. The latter assumption is motivated chiefly by the work of Edelberg (2003) discussed above.

The exercises in Athreya (2004) suggest that the elimination of stigma will produce increases in bankruptcy on the order observed in the data but will also result in sharply lower debt loads carried by households, a situation distinctly at odds with the data. However, the experiment of lowering transactions costs is able to match not only the observed increase in filing rates, but also the increases in debt/income ratios observed in U.S. data. The key intuition driving these results is first that in a low-stigma environment, borrowers will have few incentives to repay loans. To the extent that lenders recognize this, the amount lenders will be willing to lend will decrease. In the extreme case where there is neither stigma nor the possibility of a bad credit rating following a bankruptcy, unsecured lending cannot occur, as households will happily take any loan offered to them and then promptly default. By contrast, to the extent that Edelberg (2003) is accurate, credit "supply" (i.e., the willingness to lend for a given interest rate) should have expanded for purely technological reasons, allowing for the simultaneous increase in debt and filing rates observed in the data.

Even with lowered stigma, the expansionary pressure on credit supply arising from cheaper intermediation is present, though to a more limited degree. That is, stigma may indeed have fallen, only to be overwhelmed by the opposing force of technological innovation. Therefore, it may still be possible to partially reconcile reductions in stigma with observed data. In sum:

Falling stigma is not a convincing explanation for the recent rise in bankruptcy rates.

Bankruptcy and its Relation to Other Forms of Insurance

As documented by Fisher (2003), and Shepard (1984), a large portion of filers report receiving publicly funded transfers in the year in which they filed for bankruptcy. Maybe most tellingly, Sullivan et al. (2000) find that more than two-thirds of all bankruptcy filers report experiencing an income disruption near the time of filing.

The empirical work of Sullivan et al. (2002) argues that not only are many of those in bankruptcy receiving unemployment insurance, but also that a disproportionate share of those receiving unemployment insurance also file for bankruptcy. Specifically, Sullivan et al. (2002) find that the unemployment rate among bankruptcy filers is between three to four times the national average.[7] The bankruptcy rate among the unemployed is also much higher at roughly 3.5 percent, or four times higher than the overall population rate (approximately 1 percent annually in recent years). Data from the Panel Study of Income Dynamics show that 12 percent of bankruptcy filers in 1995 lost their jobs between 1994–1995, as opposed to just 2.15 percent of non-filers. Bermant and Flynn (2002) argue that bankruptcy filers also have shorter job tenure than non-filers whereby job tenure at the median for the bankrupt population is only two years, less than half of the 4.7 years of tenure for the non-bankrupt population. Sullivan et al. (2000) find that more than two-thirds of all households that file for bankruptcy report job-related income disruptions.

The quantitative evaluation of the effects of this form of insurance on behavior relating to other forms is new. The work of Livshits et al. (2003) compares bankruptcy in the United States and Germany, taking as given the structure of public insurance policies in each country. The main finding is that "fresh-start" bankruptcy is far less desirable in the presence of the comprehensive public insurance present in Germany. Athreya and Simpson (forthcoming) study an environment in which the public insurance system, including programs like the U.S. unemployment insurance system and welfare, coexist with the more implicit insurance that unsecured debt with bankruptcy may provide. Unlike prior work, Athreya and Simpson (forthcoming) not only allow more generous insurance to affect credit markets, but also to allow credit markets to affect behavior in public insurance programs. They find that the U.S. bankruptcy code obstructs public insurance provision in the United States. More surprisingly, they find that a more comprehensive social safety net provided, for example, through publicly funded insurance, will actually encourage risk-taking and reduce job search among the unemployed to such an extent that bankruptcy filings rise. More generally, the preceding makes the following clear:

Bankruptcy should be analyzed jointly with available insurance programs, because these systems can interact in perverse ways.

4. CONCLUDING REMARKS AND FUTURE WORK

Recent work has produced substantial progress in revealing the conditions prevailing at, before, and after bankruptcy. Such work has also revealed that interstate variation in rules, particularly exemptions, matter for household

[7] Bermant, Flynn, and Bakewell (2002) estimate a rate of 19 percent in 2002 from self-reported unemployment data. Athreya and Simpson (2004) find a rate of 16 percent from PSID data.

decisions and for the availability and distribution of unsecured credit. The findings established by careful empirical work has proved critical for quantitative equilibrium work that aims to account for the observations. The latter has reached some provisional conclusions.

- *Under income processes that allow for large shocks to net worth, bankruptcy can play a role in improving welfare. Without large shocks, however, bankruptcy is difficult to justify.*

- *As a quantitative matter, moral hazard needs to be taken seriously in evaluating bankruptcy provisions.*

- *Exemptions tend to distribute credit away from the asset-poor to the asset-rich but do not appear to lower welfare.*

- *Falling stigma is not a convincing explanation for the recent rise in bankruptcy rates.*

- *Bankruptcy should be analyzed jointly with available insurance programs, because these systems can interact in perverse ways.*

Despite the progress evident from the work done so far, many interesting questions remain, the resolution of which is critical for forming a definitive view of the role for personal default. From a theoretical perspective, perhaps the most useful work will be to endogenize fully the exclusion that seems to affect those who have filed for bankruptcy. Specifically, in a competitive setting, exclusion can only take place if, after a bankruptcy filing, it is optimizing to restrict lending to such a borrower. Several mutually nonexclusive possibilities arise. First, and perhaps most naturally, bankruptcy may reveal something more "permanent" about a filer. The most obvious is that a bankruptcy was triggered by a persistent shock to household earnings-generating capacity. In such an event, a household may have a very low ability or willingness to service debt. However, the desire of a household to smooth intertemporally may also be very limited. In this case, observing that a household does not borrow following bankruptcy does not pin down why it does not borrow. The problem of rationalizing exclusion may in part stem from the inability of credit market data and income data to provide high quality targets that an equilibrium model must match. Recent progress on this dimension is the work of Yue (2004), who allows for a limited form of "renegotiation" in debts between creditors and borrowers.

A second open issue is the role of stigma and an assessment of its importance in determining outcomes. While it seems unconvincing to argue that stigma has fallen, it seems quite plausible that some stigma or societal disapproval exists for bankruptcy. However, there is not a single, universally accepted measure of this. Nonetheless, knowing the extent to which shame and stigma matter is necessary for making accurate statements about the quantitative usefulness of bankruptcy to improve welfare.

A third issue is the role of bankruptcy in exacerbating moral hazard in insurance programs such as public unemployment assistance, as well as the role of the latter in encouraging bankruptcy. More work along the lines of Athreya and Simpson (forthcoming) will be useful in determining the extent to which bankruptcy and other insurance schemes (both private and public), confound each other.

A fourth issue based on the arguments above, bodes the question, why is default, regardless of whether it is legally mandated or privately contracted, useful? In the United States, however, the Constitution preserves the right of the federal government to promulgate uniform nationwide bankruptcy law. It is primarily the legal provision for bankruptcy that raises the issue of whether it improves welfare or not.

A final issue is to reconcile the two possible rationales for market incompleteness. Namely, is insurance incomplete precisely because of limited commitment arising from bankruptcy, or because of standard complications arising from asymmetric information? Future work that is able to deal effectively with these issues will greatly advance our understanding on the desirability of personal bankruptcy.

REFERENCES

Aiyagari, S. R. 1994. "Uninsured Idiosyncratic Risk and Aggregate Saving." *Quarterly Journal of Economics* 109 (3): 659–84.

Athreya, Kartik B. 2002. "Welfare Implications of the Bankruptcy Reform Act of 1999." *Journal of Monetary Economics* 49 (8): 1567–95.

_______________. 2004. "Shame As It Ever Was: Stigma and Personal Bankruptcy." *Federal Reserve Bank of Richmond Economic Quarterly* 90 (Spring): 1–19.

_______________. Forthcoming. "Fresh Start or Head Start? Uniform Bankruptcy Exemptions and Welfare." *Journal of Economic Dynamics and Control.*

_______________, and Nicole B. Simpson. Forthcoming. "Unsecured Debt with Public Insurance: From Bad to Worse." *Journal of Monetary Economics*.

Baird, Douglas G. 2001. *Elements of Bankruptcy*, 3rd Edition. New York: Foundation Press.

Bermant, Gordon, and Ed Flynn.. 2002. "Just Recently Hired: Job Tenure Among No-asset Chapter 7 Debtors." *American Bankruptcy Institute Journal*. May.

Chatterjee, Satyajit, Dean Corbae, Makoto Nakajima, and Jose-Victor Rios-Rull. 2002. "A Quantitative Theory of Unsecured Consumer Credit with Risk of Default." Federal Reserve Bank of Philadelphia: Working Paper 02-6.

Congressional Budget Office. 2000. "Personal Bankruptcy: A Literature Review." Available at http://www.cbo.gov/ftpdocs/24xx/doc2421/Bankruptcy.pdf. (Accessed 25 March 2005).

Deaton, Angus M., and Christina Paxson. 1994. "Intertemporal Choice and Inequality." *Journal of Political Economy* 102 (3): 437–67.

Dubey, Pradeep, John Geanakoplos, and Martin Shubik. 2005. "Default and Punishment in General Equilibrium." *Econometrica* 73 (1): 1–38.

Fay, Scott, Erik Hurst, and Michelle White. 2002. "The Household Bankruptcy Decision." *American Economic Review* 92 (3) 706–18.

Edelberg, W. 2003. "Risk-based Pricing of Interest Rates in Household Loan Markets." Federal Reserve Board Working Paper 2003-62. Available at http://www.federalreserve.gov/pubs/feds/2003/200362/200362pap.pdf (accessed 17 March 2005).

Elul, R., and N. Subramanian. 2002. "Forum-shopping and Personal Bankruptcy." *Journal of Financial Services Research* 21 (3): 233–55.

Fisher, Jonathan D. 2003. "The Effect of Government Unemployment Benefits, Welfare Benefits, and Other Income on Personal Bankruptcy." Manuscript, Bureau of Labor Statistics .

Grant, C., 2003. "Evidence on the Effect of U.S. Consumer Bankruptcy Exemption." Mimeo, European University Institute.

Gropp, R., J. K. Scholz, M. J. White. 1997. "Personal Bankruptcy and Credit Supply and Demand." *Quarterly Journal of Economics* 112 (1): 217–51.

Gross, David B., and Nicholas Souleles. 2002. "Explaining the Increase in Bankruptcy and Delinquency: Stigma vs. Risk Composition." University of Pennsylvania, Wharton Financial Institutions Center Working Paper 98-28-B.

Huggett, Mark. 1993. "The Risk-Free Rate in Heterogeneous-Agent Incomplete-Insurance Economies." *Journal of Economic Dynamics and Control* 17 (5–6): 953–69.

Kocherlakota, Narayana. 1996. "Implications of Efficient Risk Sharing without Commitment." *Review of Economic Studies* 63 (4): 595–609.

Li, Wenli, and Pierre-Daniel Sarte. Forthcoming. "The Macroeconomics of U.S. Consumer Bankruptcy Choice: Chapter 7 or Chapter 13?" *Journal of Monetary Economics*.

Livshits, Igor, James MacGee, and Michele Tertilt. 2003. "Consumer Bankruptcy: A Fresh Start." Federal Reserve Bank of Minneapolis, Working Paper 617.

Ljungqvist, Lars, and Thomas J. Sargent. 2000. *Recursive Macroeconomic Theory*. Cambridge, Mass.: MIT Press.

Mateos-Planas, Xavier, and Giulio Seccia. 2004. "Welfare Implications of Endogenous Credit Limits with Bankruptcy." Mimeo, Department of Economics, University of Southampton, UK.

Moss, David A., and Gibbs Johnson. 1999. "The Rise of Consumer Bankruptcy: Evolution, Revolution, or Both?" *American Bankruptcy Law Journal* 73 (2): 311–51.

Obstfeld, Maurice, and Kenneth Rogoff. 1996. *Foundations of International Macroeconomics*. Cambridge, Mass.: MIT Press.

Pavan, M. 2003. "Consumer Durables and Risky Borrowing: The Effects of Bankruptcy Protection." Mimeo, Boston College.

Shepard, Lawrence. 1984. "Personal Failures and the Bankruptcy Reform Act of 1978." *Journal of Law and Economics* 27 (2): 419–37.

Sullivan, Teresa A., Elizabeth Warren, and Jay Lawrence Westbrook. 1989. *As We Forgive Our Debtors : Bankruptcy and Consumer Credit in America*. New York: Oxford University Press.

______________. 2000. *The Fragile Middle Class: Americans in Debt.* New Haven: Yale University Press.

Townsend, R. M. 1979. "Optimal Contracts and Competitive Markets with Costly State Verification." *Journal of Economic Theory* 21 (October): 265–93.

Yue, Vivian Z. 2004. "Sovereign Default and Debt Renegotiation." Mimeo, University of Pennsylvania.

Zame, William R. 1993. "Efficiency and the Role of Default When Security Markets Are Incomplete." *American Economic Review* 83 (5) 1142–64.

Zha, Tao. 2001. "Bankruptcy Law, Capital Allocation, and Aggregate Effects: A Dynamic Heterogeneous Agent Model with Incomplete Markets." *Annals of Economics and Finance 2* (November): 379–400.