

Zero Inflation and the Friedman Rule: A Welfare Comparison

Alexander L. Wolman

A distinct trend in recent years has been for central banks to emphasize low and stable inflation as a primary goal. In many cases zero inflation—or price stability—is promoted as the ultimate long-run goal (Federal Reserve Bank of Kansas City 1996). Economic theory also stresses the benefits of low inflation. However, in contrast to the current fashion among central banks, one of the most famous—and robust—results in monetary theory is that the optimal rate of inflation is *negative*: in many economic models in which money plays a role, welfare is maximized when the inflation rate is low enough so that the nominal interest rate is zero. Central bankers are certainly aware of this result, yet they never seriously advocate a long-run policy of deflation (negative inflation).

How much welfare is lost from a zero inflation policy as opposed to an optimal deflation policy? As shown below, the shape of the economy's money demand function with respect to nominal interest rates holds the key to answering the question. Lucas (1994) argues for a specification where real balances increase toward infinity as the nominal interest rate approaches zero. He finds that zero inflation is not much of an improvement over moderate inflation but that optimal deflation offers sizable benefits. The analysis in this article supports a different conclusion: reducing inflation from a moderate level to zero entails substantial welfare benefits, and the additional benefit achieved by optimal deflation is small. My analysis is based on estimating a general

■ This article is based on the third essay in my 1996 doctoral dissertation at the University of Virginia. I would like to thank Robert King, my dissertation advisor, for his support. Thanks also to Michael Dotsey, Robert Hetzel, Andreas Hornstein, and Thomas Humphrey for their comments. The views expressed here are the author's and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

money demand function that nests the one preferred by Lucas. The estimates imply a *satiation* level of real balances, which proves to be important for the comparison of zero inflation and optimal deflation.¹

The original analysis of the relationship between money demand and the welfare cost of inflation is credited to Bailey (1956). I review both Bailey's analysis and that of Friedman (1969), whose "Friedman rule" is the famous result previously mentioned. I then describe informally Lucas's (1994) recent work on quantifying the costs of deviating from the Friedman rule. Whereas Lucas's work is guided by inventory theory, my own estimates follow from a broader interpretation of the transactions-time approach to money demand. I use these estimates for welfare analysis similar to Lucas's. Although the analysis suggests that the Friedman rule may not offer much of a benefit in comparison to zero inflation, it does not explain why central banks do not choose to pursue deflation. I thus point out several channels absent from my analysis through which inflation may have welfare effects. These additional channels may help to explain why central banks seem content to shoot for zero inflation.

1. MONEY DEMAND AND THE WELFARE COST OF INFLATION

Bailey (1956) showed how a money demand relationship could be used to derive estimates of the welfare cost of inflation. He assumed a money demand function that gave real balances (M/P , where M is the nominal quantity of money and P is the price level) as a function of the nominal interest rate (R) and made a consumer surplus argument: just as the area under the demand curve for any good measures the total private benefits of consuming that good, so the area under a money demand curve represents the private benefit of holding money. At a nominal interest rate of 5 percent, since people are willingly giving up 5 cents per year per dollar of money held, the marginal benefit of holding the last dollar must be 5 cents per year. Similarly, at a nominal interest rate of zero, people are not giving up any interest payments to hold money, so the marginal benefit of holding the last dollar must be zero. At a social optimum, the marginal benefit to society of holding money should equal the marginal cost to society of producing money. With the reasonable simplifying assumption that the cost to society of producing money is zero, the optimal nominal interest rate is zero.² In a steady state the nominal interest rate is approximately equal to the real interest rate plus the inflation rate, so optimal policy, commonly known as the Friedman rule, involves deflation at a rate equal to the real interest rate.

¹ Chadha, Haldane, and Janssen (1997) have performed an analysis similar to this article using U.K. data. They emphasize a distinction between short-run and long-run money demand.

² Lacker (1996) reports manufacturing and operating costs for coin and currency of approximately 0.2 percent of face value.

With a nominal interest rate of zero as the optimal policy, it is possible to measure the cost of any inflation rate for a particular money demand function. Simply measure the area under the inverse money demand curve between the real balances corresponding to the Friedman rule and the real balances corresponding to the nominal interest rate in question.³ That is, add up all of the marginal benefits that are foregone by following a suboptimal policy; those marginal benefits are measured by the nominal interest rate (the inverse money demand function) at each level of real balances.⁴ At this point the term “cost of inflation” may seem misleading; according to the theory sketched above, it would be more appropriate to use the term “cost of positive nominal interest rates.” Since the former term is so widely used, however, I will stick with it.

Particular theories of money may imply more complicated money demand relationships than the one assumed by Bailey; for example, the analysis in Section 3 will involve consumption and the real wage as arguments in the money demand function. However, it is still the case that the Friedman rule is optimal, and holding consumption and the real wage constant, the area under the inverse money demand curve still provides an approximate measure of the direct cost of inflation.⁵

While the optimality of the Friedman rule holds as long as real balances are a decreasing function of the nominal interest rate (subject to the caveats in Section 5), the welfare costs of inflation can vary with the money demand function in two ways. First, the overall benefit of reducing inflation from, say, 10 percent to the Friedman rule can vary. Second, the apportionment of that benefit may vary, in the following sense. According to one money demand function, reducing inflation from 10 percent to zero may generate 99 percent of the total welfare benefit, with the remaining reduction to the Friedman rule adding essentially nothing. Another function could reverse this; reducing inflation from 10 percent to zero might generate only 1 percent of the total welfare benefit, with the remaining reduction to the Friedman rule being crucial for generating any significant benefits. This article is concerned mainly with the latter issue.

Lucas (1994) contrasts the welfare implications of two particular money demand functions, both of which specify the ratio of real balances to real

³ The standard money demand curve expresses real balances as a function of nominal interest rates, whereas the inverse money demand curve inverts this relationship to express nominal interest rates as a function of real balances.

⁴ This measure of the cost of inflation does not take into account indirect effects of inflation, as will be explained in Section 4. I thus refer to the area under the money demand curve as a measure of the direct cost of inflation.

⁵ If the money demand relationship involves variables other than the nominal interest rate, the area under the inverse money demand curve ($R(m)$) only approximates the direct cost of inflation, because these other variables will generally vary across different values of the nominal interest rate.

consumption as a function of the nominal interest rate. The ratio of real balances to consumption is used because the money demand functions discussed here are assumed to apply to long-run data, and in the long run real balances move roughly one for one with consumption.⁶ In the first specification, semi-log, there is a fixed relationship between the change in the nominal interest rate and the percentage change in the real balances to consumption ratio. That is, if the nominal interest rate rises from zero to 1 percent, the percent decrease in real balances/consumption is the same as if the nominal interest rate rises from 5 percent to 6 percent. In the second specification, log-log, there is a fixed relationship between the *percentage* change in the nominal interest rate and the percentage change in the real balances to consumption ratio. Thus an increase in the nominal interest rate from zero to 1 percent will cause a much larger percentage drop in real balances/consumption than an increase in the nominal interest rate from 5 percent to 6 percent. Note that if the log-log relationship is taken literally, the ratio of real balances to consumption must be infinite when the nominal interest rate is zero.

How do the two specifications compare in terms of welfare? With the log-log function, a slight increase in the nominal interest rate near zero generates a tremendous decline in the ratio of real balances to consumption. Using Bailey's (1956) reasoning, there must be a significant welfare cost of deviating just slightly from the Friedman rule. The semi-log specification generates smaller costs of slight deviations from the Friedman rule but roughly the same benefits of reducing inflation from, say, 5 percent to zero. Lucas argues that for the United States, the log-log specification fits the data more closely than the semi-log specification.⁷ Most of the benefits to reducing inflation would then accrue only if the inflation rate were made negative, as it would need to be in order to achieve the Friedman rule. In his own words, "log-log demand implies a substantial gain in moving from zero inflation to the Friedman optimal deflation rate needed to bring nominal interest rates to zero, while under semi-log demand this gain is trivial" (Lucas 1994, p. 5).

Is log-log demand an accurate characterization of the data? Lucas argues that it is more accurate than semi-log demand, but is it reasonable to restrict the search to those two alternatives? Answering these questions requires one to be explicit about a model of money demand.

⁶ Lucas refers to the ratio of real balances to income. In his model there is no investment, so consumption equals income. The model I will use does have investment, and the appropriate ratio will be real balances to consumption rather than real balances to income.

⁷ Lucas (1994, p. 3) plots semi-log and log-log functions for various interest semi-elasticities and elasticities and concludes that "the semi-log function . . . provides a description of the data that is much inferior to the log-log curve."

2. THE TRANSACTIONS-TIME APPROACH TO MONEY DEMAND

Economists have developed a wide range of models of money, none of them entirely satisfactory. The models that are most appealing in terms of their microfoundations—that is, their descriptions of the obstacles that individuals overcome by holding money—tend to be ill-suited to quantification (e.g., estimating the welfare cost of inflation in the United States). An example is the search-theoretic class of models developed by Kiyotaki and Wright (1989).⁸ On the other hand, those models that *are* easiest to quantify do not convincingly describe the obstacles that cause individuals to hold money. Examples include the money-in-the-utility function and cash-in-advance approaches (Sidrauski 1967 and Lucas and Stokey 1983, respectively). A middle ground is the transactions-time approach, developed by McCallum (1983) and McCallum and Goodfriend (1987). Their fundamental assumption is that consumption requires time spent shopping (or transacting), and transactions time may be decreased by holding a greater quantity of real balances. The analysis in this article will be conducted in the transactions-time framework.

Denoting transactions time in period t by h_t , and the transactions-time function by $h(c, m)$, the assumptions that transactions time is increasing in consumption and decreasing in real balances mean that $\partial h/\partial c > 0$ and $\partial h/\partial m < 0$. I make the further assumption that the function is homogeneous of degree zero in c and m : if c and m increase or decrease by the same percentage, then transactions time is unchanged. It follows that only the ratio of m to c matters: $h_t = h(m_t/c_t)$. Lucas (1994) shows that the transactions-time approach can be explicitly linked to earlier inventory-theoretic models of money demand developed by Baumol (1952) and Tobin (1956). The simplest inventory-theoretic model corresponds to the transactions-time technology,

$$h(m_t/c_t) = \kappa \cdot (m_t/c_t)^{-1}, \quad (1)$$

where κ can be interpreted as a fixed cost of replenishing money holdings.⁹ More complicated inventory-theoretic approaches can be shown to imply similar $h(\cdot)$ functions, with the difference being that m/c would be raised to some power less than -1 :

$$h(m_t, c_t) = \kappa \cdot (m_t/c_t)^{-1/\gamma}, \gamma \in (0, 1). \quad (2)$$

See Lucas (1994).

⁸ This is not to rule out the possibility that in the future, search-based models will be useful for quantitative exercises.

⁹ While McCallum and Goodfriend interpreted $h(\cdot)$ in terms of shopping time, Lucas interpreted it as going-to-the-bank time.

The inventory-theoretic interpretation imposes strong restrictions on the form of the transactions-time technology and hence, as I will describe below, on the form of the money demand function. Specifically, for the transactions-time technology, it implies that no matter how high the ratio of real balances to consumption, there is still some additional benefit to increasing that ratio further. Lucas (1994, p. 16) defends this implication as follows: “Managing an inventory always requires *some* time, and a larger average stock must always reduce this time requirement, no matter how small it is.” One cannot argue with this statement, according to a narrow interpretation of what it means to manage an inventory. However, holding a higher inventory of real balances also requires increased resources to protect the inventory, a point made by Friedman (1969, p. 17), who described a shopkeeper hiring guards to “protect his cash hoard.”

Given an arbitrary transactions-time technology, the associated money demand function can be derived by specifying some additional features of the economic environment. Assume that individuals face a budget constraint,

$$P_t c_t + M_t + \frac{B_t}{1 + R_t} = M_{t-1} + B_{t-1} + P_t w_t n_t + D_t, \quad (3)$$

and a time constraint,

$$n_t + l_t + h_t = 1, \quad (4)$$

where P_t is the price level, M_t is nominal money balances ($m_t p_t$), B_t is holdings of one-period nominal zero-coupon bonds maturing at $t + 1$, R_t is the interest rate on bonds, w_t is the real wage, n_t is the fraction of time spent working, D_t is dividend payments from firms, l_t is the fraction of time spent as leisure, and h_t is the fraction of time spent carrying out transactions. In a given period, individuals’ sources of funds are the money balances with which they enter the period, the bonds they redeem, the wage income they earn, and the dividends they receive from firms. These sources fund current consumption and money balances and bonds to carry over into the next period.

Deriving the money demand function requires knowing what it means for an individual to hold an optimal quantity of real balances. Optimal behavior involves balancing marginal benefit and marginal cost. What are the marginal benefit and marginal cost of holding money? From Section 1, the marginal cost of an additional dollar is the interest foregone in the next period (R_t); the marginal benefit of an additional dollar is the decrease in transactions time that it brings about. This decrease in transactions time is $-h'(m_t/c_t) \cdot \frac{1}{P_t \cdot c_t}$, and the extra time can be spent in the labor market earning the nominal wage ($P_t \cdot w_t$). Since marginal cost is measured as of the subsequent period, marginal benefit needs to be adjusted correspondingly: current period labor earnings can be invested in the bond market, so their value tomorrow is $\left(-P_t \cdot w_t \cdot (1 + R_t) \cdot \frac{1}{P_t \cdot c_t} \cdot h'(m_t/c_t)\right)$.

Equating marginal cost and marginal benefit implies

$$-h'(m_t/c_t) = \frac{R_t}{1 + R_t} \cdot \frac{c_t}{w_t}, \quad (5)$$

which can be used to confirm the Friedman rule result: at a nominal interest rate of zero, money holdings are chosen so that the marginal benefit of an additional unit of money is zero.

Under the inventory-theoretic interpretation, as mentioned earlier, the marginal benefit of an additional unit of money is never zero. Combining (5) with the specification in (2), the strictly positive marginal benefit of additional real balances corresponds to infinite real balances at the Friedman rule ($R = 0$):

$$\frac{\kappa}{\gamma} \cdot (m_t/c_t)^{-1-1/\gamma} = \frac{R_t}{1 + R_t} \cdot \frac{c_t}{w_t}, \gamma \in (0, 1]. \quad (6)$$

The inventory-theoretic approach has appeal, but the implication that real balances would be infinite at the Friedman rule is extreme and argues for considering transactions-time technologies that do not share that implication. If real balances are finite at the Friedman rule, there is some quantity of real balances at which the marginal benefit of holding an additional unit of real balances is zero. That level of real balances—if it exists—will be referred to as the *satiation* level. A key proposition, namely that the welfare gains from low nominal interest rates are concentrated near the Friedman rule, depends crucially on the assumption of no satiation level; the log-log money demand function does not have satiation, whereas the semi-log function does.

The log-log function is roughly consistent with inventory theory: assuming that c and w are constant, and noting that $R_t/(1 + R_t) \approx R_t$, (6) yields a nearly linear relationship between the log of real balances and the log of the nominal interest rate. In contrast, the semi-log function is inconsistent with inventory theory, as it posits a linear relationship between the log of real balances and the *level* of the nominal interest rate. Thus Lucas's purely empirical argument favoring the log-log specification over semi-log is strengthened by his theoretical argument favoring the inventory approach. However, inventory-theoretic models do not offer the only alternative to semi-log money demand. And the fact that the inventory approach implies infinite real balances at a zero nominal interest rate suggests searching across a wider class of models. In the next section, I present estimates of a money demand function that allows for satiation and is consistent with the basic assumptions of the transactions-time model. This function nests nonsatiation (log-log) as a special case, but for many parameter values it is not consistent with inventory theory.

3. ESTIMATES OF A GENERAL MONEY DEMAND FUNCTION

From (5), in order for a transactions-time function to be consistent with satiation, it must be that for some positive value of m/c , further increases in that ratio do not decrease transactions time. In (6), under the inventory approach, transactions time is always decreasing in m/c , so subtracting a constant from the left-hand side of (6) will yield a technology consistent with satiation. That is, if $h'(m_t/c_t) = \phi - (\kappa/\gamma) \cdot (m_t/c_t)^{-1-1/\gamma}$, with $\phi \geq 0$, then the implied transactions-time technology allows for satiation. Since it will be convenient below to specify the parameters in a slightly different way, I define $\nu \equiv -\gamma/(1+\gamma)$, and $A \equiv (\kappa/\gamma)^{-\gamma/(1+\gamma)}$, so that $h'(m_t/c_t) = \phi - A^{-1/\nu} \cdot (m_t/c_t)^{1/\nu}$, with $\nu < 0$, $A > 0$. The technology can be found by integrating the previous expression:

$$h(m_t/c_t) = \phi \cdot (m_t/c_t) - \frac{\nu}{1+\nu} A^{-1/\nu} \cdot (m_t/c_t)^{\frac{1+\nu}{\nu}} + \Omega, \text{ for } m_t/c_t < A \cdot \phi^\nu, \quad (7)$$

$$h(m_t/c_t) = \Omega, \text{ for } m_t/c_t \geq A \cdot \phi^\nu,$$

where Ω is a nonnegative constant that represents the minimum possible transactions time. This function is decreasing in m_t/c_t as long as m_t/c_t is less than $A \cdot \phi^\nu$, and the satiation level of real balances is given by $(m/c)_s = A \cdot \phi^\nu$. If $\phi = 0$, then there is no satiation level, and the function is consistent with inventory theory. The implicit money demand function is given by

$$A^{-1/\nu} \cdot (m_t/c_t)^{1/\nu} - \phi = \frac{R_t}{1+R_t} \cdot \frac{c_t}{w_t}, \quad (8)$$

which can be rewritten to yield an explicit money demand function:

$$m_t/c_t = A \cdot \left(\frac{R_t}{1+R_t} \cdot \frac{c_t}{w_t} + \phi \right)^\nu. \quad (9)$$

My strategy now is to estimate A , ϕ , and ν using (9) and to test the hypothesis that there is no satiation level of real balances ($\phi = 0$). The theory as presented thus far suggests that (9) should hold exactly. Of course it does not; I choose to model the error term as additive, but the estimation results do not change significantly if the error is assumed to be multiplicative. The data, which are from the United States for the period 1915 to 1992, are described in the appendix.

Although four separate variables enter (9), for estimation purposes it is simplest to define the two composite variables, $y_t \equiv m_t/c_t$ and $x_t \equiv [R_t/(1+R_t)] \times [c_t/w_t]$. Then the estimation equation is

$$y_t = A \cdot (x_t + \phi)^\nu + \varepsilon_t. \quad (10)$$

Figure 1 displays a plot of y_t versus x_t . Estimates of A , ν , and ϕ are found by solving the following nonlinear least squares (NLS) problem:¹⁰

$$\min_{\hat{A}, \hat{\nu}, \hat{\phi}} \sum_{t=1}^T \left(y_t - \hat{A} \cdot (x_t + \hat{\phi})^{\hat{\nu}} \right)^2. \quad (11)$$

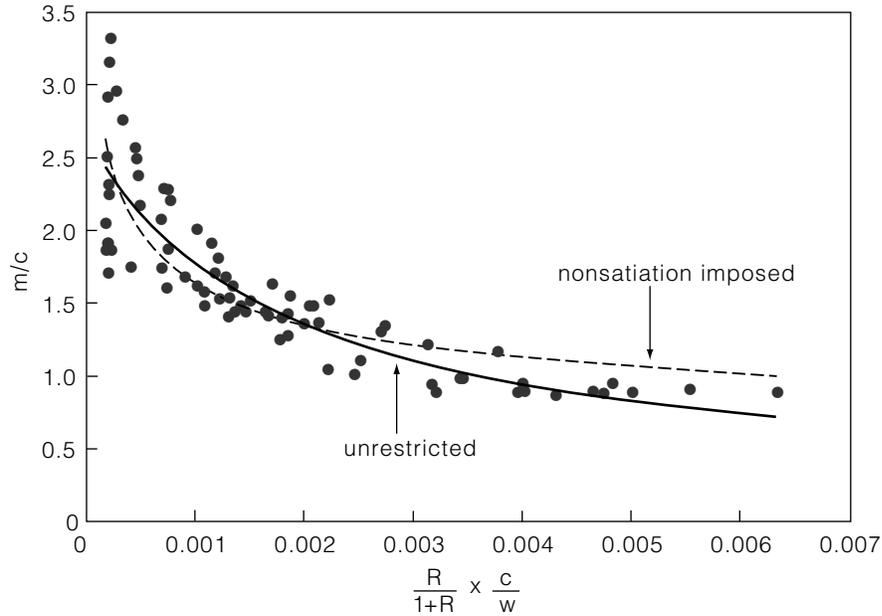
In general, the NLS estimates are consistent and asymptotically normal, as shown by Amemiya (1985, pp. 127–35); here I do not make any distributional assumptions about ε_t , the residual. Confidence intervals for the parameters were generated by bootstrapping, which allows one to construct a sampling distribution without any distributional assumptions and without relying on the accuracy of linear approximations.¹¹

Table 1 contains estimates for A , ν , and ϕ , along with centered 95 percent confidence intervals. Although the estimated value of ϕ is close to zero, the implied satiation level of m/c is fairly low, 2.674. Following Amemiya (1985, p. 136), I construct a t-test of the nonsatiation hypothesis ($\phi = 0$). The test statistic is 26.99, meaning that nonsatiation is overwhelmingly rejected. Using the sampling distribution for the parameters A , ϕ , and ν , Figure 2 plots the implied sampling distribution for the satiation level of m/c . According to the sampling distribution, 90 percent of the probability mass for the satiation ratio lies below a value of 5. However, the right-hand tail of the distribution is fat; the x-axis would need to go all the way to 46,000 to encompass 97.5 percent of the probability mass, meaning that the satiation level is imprecisely estimated. This imprecision follows from the properties of the data: the lowest nominal interest rate in the sample is 0.7 percent, and for the observations with the lowest nominal interest rates, there is substantial variation in the ratio of real balances to consumption.¹² The solid line in Figure 1 shows the fitted values.

¹⁰ The presence of consumption in the numerator of x and the denominator of y can cause the NLS estimator to be biased, as it may induce a correlation between the residual (ε_t) and x . More generally, if the residual represents a shock to the transactions-time technology, then in general equilibrium such a correlation would arise even without consumption on both sides of the estimation equation. I have investigated these problems by estimating with instrumental variables using the generalized method of moments (GMM). The GMM estimates are highly sensitive to the choice of instruments, so I report only the NLS results.

¹¹ The bootstrapping approach involves three steps. The first step is to produce the NLS estimates. The second step is to fit an AR model to the NLS residuals, producing a new set of disturbances, $\hat{\varepsilon}_t$, that are approximately white noise (an AR(2) was fit to $\hat{\varepsilon}_t$ to produce $\hat{\varepsilon}_t$). The final step is to draw randomly with replacement from the $\hat{\varepsilon}_t$, producing N new vectors, \tilde{y}_t , each of size T . For each of those new samples the parameters are estimated by nonlinear least squares. The \tilde{y}_t are generated by combining the x_t data and the random draws of $\hat{\varepsilon}_t$ with the initial parameter estimates.

¹² Working in a different money demand framework, Mulligan and Sala-i-Martin (1996) have developed a method of estimating the behavior of money demand near zero nominal interest rates. Their fundamental insight is that if there is a fixed cost of holding nonmonetary assets, the behavior of individuals who hold only monetary assets at positive nominal interest rates can yield information about aggregate money demand at a nominal interest rate of zero.

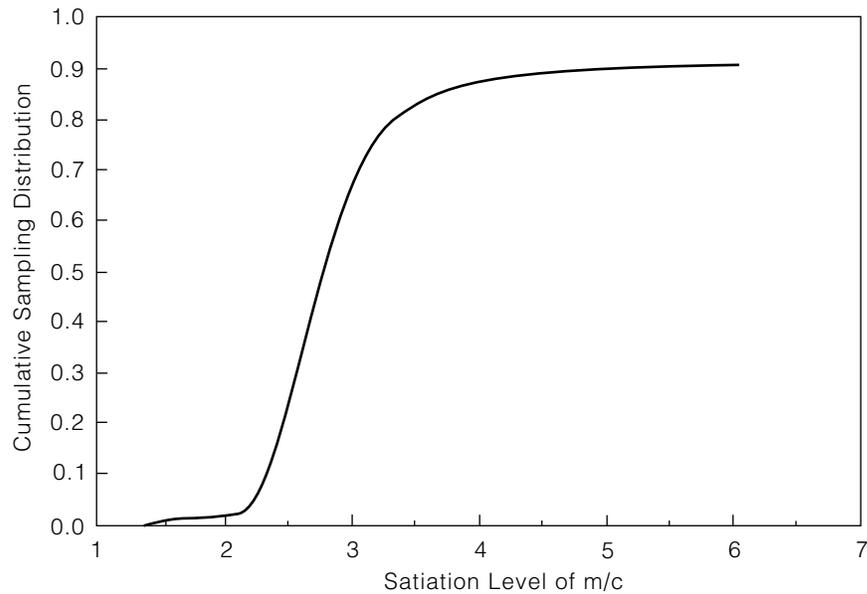
Figure 1 Data and Predicted Values**Table 1 Unrestricted Estimates and 95 Percent Confidence Intervals**

\hat{A}	$\hat{\nu}$	$\hat{\phi}$
0.01702	-0.7695	0.001399
$(6.1 \times 10^{-7}, 0.421)$	$(-3.31, -0.20335)$	$(2.5 \times 10^{-19}, 0.0131)$

For comparison purposes, I also estimated A and ν under the nonsatiation restriction. Table 2 contains the estimates, and the dashed line in Figure 1 shows the fitted values when nonsatiation is imposed. With money demand estimates in hand, we can now look at their implications for the welfare cost of inflation.

Table 2 Restricted Estimates: Nonsatiation

\hat{A}	$\hat{\nu}$
0.2526	-0.2699

Figure 2 Distribution of Estimated Satiation Level

4. WELFARE ANALYSIS

By specifying a general equilibrium model, I can use the above estimates of the transactions-time technology to compute the exact welfare cost of inflation. I use a standard real business cycle model, as in Prescott (1986) or King, Plosser, and Rebelo (1988), augmented by the transactions-time money demand specification to answer the following question: in a world of constant 5 percent inflation, how much income would an individual willingly forfeit (or require) in order to live in a world with some lower (or higher) constant inflation rate?¹³

The economy consists of a representative individual who chooses consumption and money balances and leisure to maximize lifetime utility:

$$E_t \sum_{j=0}^{\infty} \beta^j u(c_{t+j}, l_{t+j}),$$

¹³ For the purpose of computing this welfare measure, I define *full income* as the sum of consumption and $w \cdot l$, where w is the real wage and l is leisure. The computation holds the real wage constant at its benchmark level. That is, what amount of additional full income at the old real wage would give the individual the same utility as the decrease in inflation under consideration?

where $u(c, l) = \ln(c) + \psi \ln(l)$. This maximization is subject to the budget constraint (3), the transactions-time technology, and the time constraint (4). Optimal choices of consumption, leisure, bond holdings, and money holdings imply

$$u_c(c_t, l_t) = \lambda_t \cdot P_t \cdot \left(1 + w_t h' \left(\frac{c_t}{m_t} \right) \left(\frac{1}{m_t} \right) \right), \quad (12)$$

$$u_l(c_t, l_t) = \lambda_t \cdot w_t \cdot P_t, \quad (13)$$

and

$$1 + R_t = E_t \frac{\lambda_t}{\beta \lambda_{t+1}}, \quad (14)$$

as well as the money demand relationship (5). In these expressions λ_t is the shadow price of nominal wealth—the multiplier on (3). Since consumption requires a time expenditure, there is a wedge between the marginal utility of consumption and the marginal utility of wealth in (12). That wedge, $\lambda_t \cdot w_t \cdot P_t \cdot h' \left(\frac{c_t}{m_t} \right) \cdot \left(\frac{1}{m_t} \right)$, is the value in utility terms of the marginal transactions time associated with an additional unit of consumption. The efficiency condition for leisure, (13), sets the marginal utility of leisure equal to the marginal utility of foregone earnings, and the efficiency condition for bond holding, (14), describes the equivalence between having \$1 of wealth today and $\$(1 + R)$ of wealth tomorrow. An additional equation defines transactions time as (7).

Firms produce the economy's single good using capital, which they own, and labor, which they hire on a period-by-period basis, according to a constant returns to scale production function,

$$y_t = a_t f(k_t, g^t n_t), \quad (15)$$

where y_t is output, a_t is a random productivity factor, k_t is the capital stock, and g is the exogenous growth rate of labor-augmenting technical progress. In a steady state, the exogenous technical progress will mean that output, consumption, real balances, the capital stock, investment, and the real wage will also grow at rate g . Capital accumulates according to

$$k_{t+1} = k_t \cdot (1 - \delta) + i_t, \quad (16)$$

where i_t is investment and δ is the depreciation rate. Since firms own the capital stock, they earn rents in equilibrium; those rents are paid out as dividends to individuals, who own the firms. Firms maximize the expected discounted stream of future profits—all of which are paid out as dividends—where the discount rate for period $t + j$ is the consumer's marginal rate of substitution between a dollar of wealth in periods t and $t + j$:

$$V_t = \text{Max } E_t \sum_{j=0}^{\infty} \beta^j \cdot \frac{\lambda_{t+j}}{\lambda_t} \cdot (P_t \cdot a_t \cdot f(k_t, g^t n_t) - w_t \cdot P_t \cdot n_t - P_t \cdot i_t).$$

This maximization is subject to (15) and (16). Thus the firm's first-order condition with respect to next period's capital stock is

$$P_t = \beta \cdot E_t \frac{\lambda_{t+1}}{\lambda_t} \cdot (P_{t+1} \cdot a_{t+1} \cdot f_k(k_{t+1}, g^{t+1} n_{t+1}) + P_{t+1} \cdot (1 - \delta)). \quad (17)$$

According to (17), the decrease in current-period profit that results from a marginal increase in investment should be exactly offset by the increase in future profits associated with a higher capital stock next period. Optimal choice of labor input implies that the real wage equals the marginal product of labor:

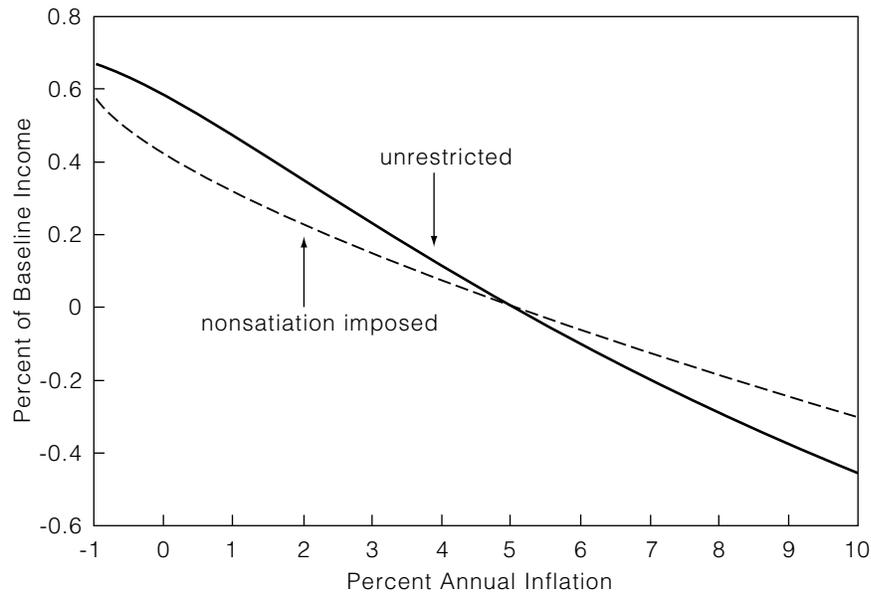
$$a_t \frac{\partial f(k_t, g^t n_t)}{\partial n_t} = w_t. \quad (18)$$

This completes the description of the model economy.

The benchmark economy has 5 percent annual inflation, and the other parameters are chosen as follows. The real growth rate (g) is 3 percent annually. The production function is Cobb-Douglas with labor's share equal to $2/3$, and the depreciation rate, δ , is 0.025. The preference parameters ψ and β are set so that the real interest rate is 6.5 percent annually and steady-state hours worked are 20 percent of the time endowment.¹⁴ All of the above values are within the normal range chosen in the real business cycle literature. Given the estimated parameters of the transactions-time technology, the constant of integration (Ω) is chosen so that steady-state transactions time is 2 percent of the time endowment, consistent with the data presented by Andreyenkov, Patrushev, and Robinson (1989). The values of the parameters ψ , β , and Ω , as well as the remaining endogenous variables, are found by solving for a deterministic steady state of the system of equations given by (12)–(18), (4), (7), and (9). To compute the welfare measure, the inflation rate alone is varied, and the new steady state is computed at each desired inflation rate.

Figure 3 plots the quantity of full income (defined in footnote 13), as a percentage of its benchmark level, that individuals would be willing to forego (would require) to live in a lower (higher) inflation world. The solid line represents the unrestricted estimated money demand specification, and the dashed line represents the restricted estimates that impose nonsatiation. Both specifications imply that if inflation were reduced to the Friedman rule from a 5 percent annual rate, for individuals in the model economy it would be as if their full

¹⁴The real interest rate, r , is equal to $(1 + R)/(1 + \pi)$, where π is the inflation rate. As in King and Wolman (1996), I assume that the risk-free real interest rate relevant for calculating the opportunity cost of money holding is 1 percent annually, whereas the real rate that implicitly enters (15) and (18) is 6.5 percent. The risk premium is not modeled explicitly. In practice this means that there are two real interest rates in the model, but since both of them are "known," no equations or unknowns are added to the steady-state computation. See below for a discussion of the implications of this ad hoc approach.

Figure 3 Welfare Compared to Baseline of 5 Percent Inflation

income had increased by about 0.6 percent.¹⁵ However, the apportionment of these benefits differs in the two cases. Under nonsatiation, less than 3/4 of this benefit can be achieved with zero inflation, whereas the unrestricted money demand specification implies that almost 9/10 of the benefit can be achieved with zero inflation. While I have argued that nonsatiation is an implausible assumption, one should keep in mind that the satiation level was imprecisely estimated. To the extent that one believes the actual satiation level is higher than it was estimated in Section 3, there would be higher costs associated with zero inflation relative to the Friedman rule than are indicated by the solid line.

While the results are not sensitive to small changes in most of the model's other parameters, they are sensitive to the underlying real interest rate. I have assumed that the real return on capital is 6.5 percent annually and that the risk-free real rate of return relevant for measuring the opportunity cost of holding money is 1 percent. Since uncertainty is not explicitly incorporated into the model, this assumption is ad hoc. The assumption is made because in the United States these have been the average real returns on equity and Treasury bills, respectively. Ideally, one would explicitly model the banking system and

¹⁵ The magnitude of these welfare benefits is similar to the magnitudes reported in Lucas (1994).

thus endogenize the spread between risky and (nominally) riskless returns. The assumption of a 1 percent riskless real rate is important for the magnitude of the welfare cost of inflation. If that rate were instead assumed to be 6.5 percent, then the Friedman rule would involve a 6.5 percent deflation rate instead of a 1 percent deflation rate. From Figure 3, this would imply a significantly larger benefit to achieving the Friedman rule. However, the behavior of money demand at the Friedman rule—that is, whether or not there is satiation—would remain important regardless of the assumed value for the real rate.

There is a long tradition, discussed earlier, of measuring the cost of inflation by the area under a money demand curve. Here that calculation would have yielded curves almost identical in shape to those in Figure 3. However, the area calculation would describe time saving only, without accounting for the effect on welfare of changes in consumption. In the case of the estimates with satiation, there is roughly a 1 percent difference in the level of consumption between the Friedman rule and 5 percent inflation. I take this difference in consumption into account in Figure 3. In general, the area under the money demand curve may misstate the welfare cost of inflation because it measures only the direct effect of increases in real balances; here the direct effect is the decrease in transactions time and the indirect effects are summarized by the increase in consumption. This distinction is especially important in Dotsey and Ireland (1996), where the (endogenous) growth rate of the economy is indirectly affected by the inflation rate. The direct effect on money demand is dwarfed by the indirect effect on growth in their model. An additional reason for preferring exact welfare calculations is that the area under the money demand curve does not take into account agents' preferences and thus cannot actually be interpreted in terms of welfare.

5. OTHER EFFECTS OF INFLATION

The above analysis compares different rates of steady inflation in a model where the only welfare effects of inflation work through the demand for money. This narrow focus was chosen to highlight the importance of assumptions about the behavior of money demand at low nominal interest rates. However, in more general models, the quantitative results involving money demand may vary. Furthermore, there may be welfare effects of inflation unrelated to the demand for money. In this section, I briefly discuss some ways in which analysis of the welfare effects of inflation differs in more general models. The references I provide are meant to serve as entry points to what in each case are extensive literatures.

Much of the literature on macroeconomic models with money has involved nominal rigidities, such as sticky prices. In contrast, the model in this article has flexible prices. Sticky prices lead to effects of steady inflation that work

through other channels in addition to money demand. Models with sticky prices usually involve imperfect competition, and inflation can affect the magnitude of the distortion from imperfect competition. In King and Wolman (1996), for example, the markup of price over marginal cost—which is a distortion—varies with inflation because firms incorporate into their pricing decisions the possibility that the price they choose will remain fixed for several periods. While some have suggested that inflation can have beneficial effects on the markup (Rotemberg 1996; Benabou 1992), King and Wolman (1996) find the opposite effect, as firms choose a high markup when they set price to compensate for the deterioration that will be caused by inflation. Whether that result generalizes to a wider class of models is an open question.

A literature beginning with Phelps (1973) extends the type of analysis performed in this article by incorporating distortionary taxes. Inflation, or more properly, money creation, is a source of revenue (seigniorage) for the government. Implicitly, my analysis has assumed that this revenue can be replaced by a lump sum tax, which does not distort individual decisions. If lump sum taxes are unavailable, so that seigniorage must be replaced by a distortionary tax such as an income tax, then the optimal rate of inflation in principle could be higher than that corresponding to the Friedman rule; there would be a welfare benefit to inflation counteracting the welfare cost associated with money demand. Recent work by Chari, Christiano, and Kehoe (1996) and Correia and Teles (1997), among others, suggests that this benefit is small enough that the Friedman rule remains optimal with distortionary taxes for a wide range of money demand specifications. With satiation, however, distortionary taxes would probably make the optimal nominal interest rate positive, because with satiation the marginal welfare cost of inflation is zero at the Friedman rule.

Feldstein (1997) has emphasized another way in which inflation interacts with public finance, namely the costs of inflation that result from a nonindexed tax code. With a nonindexed tax code, inflation raises the effective tax rate on both individuals and businesses. Feldstein argues that these tax-related distortions alone cost the U.S. economy about 0.8 percent of GDP per year.

Finally, the steady-state analysis in this article leaves open the question of transitional effects of a significant decrease in the inflation rate. These transitional effects would be small in the model used here. However, models with sticky prices or other nominal rigidities may imply significant welfare costs of a transition to lower inflation, with the costs depending on such factors as how credible the disinflation is. Friedman himself stressed transitional issues: “Any decided change in the trend of prices would involve significant frictional distortion in employment and production” (1969, p. 45). This topic is currently being studied intensively; see Ball (1994a,b) and Ireland (1995) for examples of recent work. It is important to note, however, that in contrast to a one-time cost of lowering the inflation rate, the benefit of low inflation emphasized in this article accrues year after year.

6. CONCLUSIONS

At positive nominal interest rates, individuals incur an opportunity cost by holding money instead of interest-bearing securities. Since the social cost of producing money is nearly zero, there is a divergence between the private and social costs of holding money when nominal interest rates are positive. Individuals choose to equate the marginal benefit of holding money with the private cost, so positive nominal interest rates generate an inefficiency. Policy-makers, by setting the nominal interest rate at zero, and so equating private and social costs, can eliminate this inefficiency. In models where there are no other distortions, it follows that this same monetary policy is optimal from a welfare perspective. Lucas (1994) has argued that the form of the money demand function implies significant welfare losses at even very low nominal interest rates. His conclusion results from his assumption that individuals do not become satiated with real balances as the nominal rate declines toward zero. Equivalently, the marginal benefit of holding real balances is positive no matter how high are individuals' money holdings.

I have estimated the money demand function implied by a general transactions-time technology and found evidence that the marginal benefit of holding real balances declines to zero at a nominal interest rate of zero. In other words, individuals can become satiated with real balances. My conclusions regarding satiation, however, are vulnerable to the criticism that zero nominal interest rates have never occurred. Nonetheless, my results imply that the welfare cost of low nominal interest rates is small. Most of the benefits from reducing inflation below, say, 5 percent can be achieved with price stability (zero inflation), and those benefits are significant. In my model a reduction in inflation from 5 percent to zero is equivalent to an increase in consumption of 0.6 percent of output. This result helps reconcile the optimality of zero nominal interest rates with the tendency of central banks to emphasize zero inflation. Still, one wonders why central banks do not simply advocate the optimal policy. Probably the explanation involves factors such as transitional costs of disinflation. This point aside, it is easier to understand why central banks would advocate sub-optimal policy if that policy is close to being optimal.

APPENDIX

The data used to estimate (9) are annual, from 1915 to 1992. The nominal interest rate is the yield on commercial paper from the National Bureau of Economic Research (NBER) database (1915–1946) and Citibase (1947–1992).

I use nominal data for consumption, the wage rate, and the money supply; taking ratios causes the price indexes to cancel. The consumption series consists of three spliced series. From 1915 to 1929, I combine personal consumption expenditures per capita in 1929 dollars, with the deflator for the same. The former is series A25 from Kendrick, reproduced in the U.S. Commerce Department's *Long-Term Economic Growth (LTEG)*. The latter is series B64 from *LTEG*. Both are annual series. From 1930 to 1945, I combine personal consumption expenditures per capita in 1958 dollars, with the deflator for the same. The former is series A26 and the latter is series B65, both from *LTEG*, and both annual. Finally, from 1946 to 1992, I use personal consumption expenditures in current dollars, divided by population. The former is series GC, from Citibase; it is in billions of dollars and is seasonally adjusted quarterly data, which I average to create annual data. The latter is PAN (Citibase 1946–1991), with data for 1992 estimated by extrapolating the average rates of change from 1990 to 1991; population is in thousands.

As mentioned above, I use nominal wage data. Also, since the raw wage data is hourly, I multiply by the number of hours in a quarter (2,184) to get a quarterly wage. From 1915 to 1946, I “reflate” total compensation per hour at work for manufacturing production workers, using the CPI. The former is series B70 from *LTEG*; it is in 1957 dollars. The latter is m04045 from the NBER database. From 1947 to 1992, I use average hourly earnings of production workers in manufacturing, in current dollars. This is series LEHM from Citibase. Finally, since the relevant wage variables from a theoretical perspective are after-tax wages, I multiply wages by the average marginal tax rates provided by Barro and Sahasakul (1983) and updated through 1992 in the manner they describe.¹⁶

For money, from 1915 to 1970 I use the M1 series from Friedman and Schwartz (1963) and the Federal Reserve, which is reproduced as series B109 and B110 in *LTEG*. From 1970 to 1992 I use FM1 from Citibase. Both series are in billions of dollars and are deflated by the POPM population measure mentioned above. Prior to 1946, that population measure is the annual series in the Bureau of the Census's *Historical Statistics* (Series A–6–8, p. 8).

¹⁶ The conclusions reached above are unchanged if before-tax wage rates are used.

REFERENCES

- Amemiya, Takeshi. *Advanced Econometrics*. Cambridge, Mass.: Harvard University Press, 1985.
- Andreyenkov, Vladimir G., Vasily D. Patrushev, and John P. Robinson. *Rhythm of Everyday Life: How Soviet and American Citizens use Time*. Boulder, Colo.: Westview Press, 1989.
- Bailey, Martin J. "The Welfare Cost of Inflationary Finance," *Journal of Political Economy*, vol. 64 (April 1956), pp. 93–110.
- Ball, Laurence. "Credible Disinflation with Staggered Price-Setting," *American Economic Review*, vol. 84 (March 1994a), pp. 282–89.
- . "What Determines the Sacrifice Ratio," in N. Gregory Mankiw, ed., *Monetary Policy*. Chicago: The University of Chicago Press, 1994b.
- Barro, Robert J., and Chaipat Sahasakul. "Measuring the Average Marginal Tax Rate from the Individual Income Tax," *Journal of Business*, vol. 56 (October 1983), pp. 419–52.
- Baumol, William J. "The Transactions Demand for Cash: An Inventory-Theoretic Approach," *Quarterly Journal of Economics*, vol. 66 (November 1952), pp. 545–66.
- Benabou, Roland. "Inflation and Markups: Theories and Evidence from the Retail Trade Sector," *European Economic Review*, vol. 36 (April 1992), pp. 566–74.
- Bureau of the Census. *Historical Statistics of the United States, Colonial Times to 1970*. Washington: U.S. Department of Commerce, 1975.
- Chadha, Jagjit S., Andrew G. Haldane, and Norbert G. J. Janssen. "Shoe-Leather Costs Reconsidered." Manuscript. Bank of England, January 1997.
- Chari, V. V., Lawrence J. Christiano, and Patrick J. Kehoe. "Optimality of the Friedman Rule in Economies with Distorting Taxes," *Journal of Monetary Economics*, vol. 37 (April 1996), pp. 203–23.
- Correia, Isabel, and Pedro Teles. "The Optimal Inflation Tax," Discussion Paper 123. Institute for Empirical Macroeconomics, Federal Reserve Bank of Minneapolis, August 1997.
- Dotsey, Michael, and Peter N. Ireland. "The Welfare Cost of Inflation in General Equilibrium," *Journal of Monetary Economics*, vol. 37 (February 1996), pp. 29–47.
- Federal Reserve Bank of Kansas City. *Achieving Price Stability*. Kansas City: Federal Reserve Bank of Kansas City, 1996.

- Feldstein, Martin. "The Costs and Benefits of Going from Low Inflation to Price Stability," in Christina D. Romer and David H. Romer, eds., *Monetary Policy*. Chicago: University of Chicago Press, 1997.
- Friedman, Milton. "The Optimum Quantity of Money," in *The Optimum Quantity of Money, and Other Essays*. Chicago: Aldine Publishing Company, 1969.
- _____, and Anna J. Schwartz. *A Monetary History of the United States 1867–1960*. Princeton: Princeton University Press, 1963.
- Ireland, Peter N. "Optimal Disinflationary Paths," *Journal of Economic Dynamics and Control*, vol. 19 (November 1995), pp. 1429–48.
- King, Robert G., Charles I. Plosser, and Sergio T. Rebelo. "Production, Growth and Business Cycles: I. The Basic Neoclassical Model," *Journal of Monetary Economics*, vol. 21 (March/May 1988), pp. 195–232.
- King, Robert G., and Alexander L. Wolman. "Inflation Targeting in a St. Louis Model of the 21st Century," *Federal Reserve Bank of St. Louis Review*, vol. 78 (May/June 1996), pp. 83–107.
- Kiyotaki, Nobuhiro, and Randall Wright. "On Money as a Medium of Exchange," *Journal of Political Economy*, vol. 97 (August 1989), pp. 927–54.
- Lacker, Jeffrey M. "Stored Value Cards: Costly Private Substitutes for Government Currency," *Federal Reserve Bank of Richmond Economic Quarterly*, vol. 82 (Summer 1996), pp. 1–25.
- Lucas, Robert E., Jr. "On the Welfare Cost of Inflation," Working Paper 394. Stanford University: Center for Economic Policy Research, 1994.
- _____, and Nancy L. Stokey. "Optimal Fiscal and Monetary Policy in an Economy without Capital," *Journal of Monetary Economics*, vol. 12 (July 1983), pp. 55–93.
- McCallum, Bennett T. "The Role of Overlapping-Generations Models in Monetary Economics," *Carnegie-Rochester Conference Series on Public Policy*, vol. 18 (Spring 1983), pp. 9–44.
- _____, and Marvin S. Goodfriend. "Demand for Money: Theoretical Studies," in *The New Palgrave: A Dictionary of Economics*. London: Macmillan Press, 1987, reprinted in *Federal Reserve Bank of Richmond Economic Review*, vol. 74 (January/February 1988), pp. 16–24.
- Mulligan, Casey B., and Xavier X. Sala-i-Martin. "Adoption of Financial Technologies: Implications for Money Demand and Monetary Policy," *NBER Working Paper 5504*. March 1996.
- Phelps, Edmund S. "Inflation in the Theory of Public Finance," *Swedish Journal of Economics*, vol. 75 (March 1973), pp. 67–82.

Prescott, Edward C. "Theory Ahead of Business Cycle Measurement," Federal Reserve Bank of Minneapolis *Quarterly Review*, vol. 10 (Fall 1986), pp. 9–22.

Rotemberg, Julio J. "Commentary," Federal Reserve Bank of St. Louis *Review*, vol. 78 (May/June 1996), pp. 108–11.

Sidrauski, Miguel. "Rational Choice and Patterns of Growth in a Monetary Economy," *American Economic Review*, vol. 57 (May 1967), pp. 534–44.

Tobin, James. "The Interest-Elasticity of Transactions Demand for Cash," *Review of Economics and Statistics*, vol. 38 (August 1956), pp. 241–47.

U.S. Department of Commerce. *Long-Term Economic Growth, 1860–1970*. Washington: Bureau of Economic Analysis, 1973.

Group Lending and Financial Intermediation: An Example

Edward S. Prescott

Imagine a small group of people, each of whom borrows money from a financial intermediary. The intermediary does not require collateral because the borrowers are relatively poor and do not own much property. Instead, the intermediary requires group members to be jointly liable for each other's loans. That is, if a member defaults on a loan, the rest of the group is liable for the remainder of the loan. If the group does not honor this joint obligation, then the entire group is cut off from future access to credit.

The lending arrangement I just described is not fictitious. Two million villagers, most of whom are female and poor, borrowed in this way from the Grameen Bank in Bangladesh. In Bolivia, 75,000 urban entrepreneurs, roughly one-third of the banking system clientele, borrowed money via group loans from BancoSol. Even in nineteenth-century Ireland, many rural residents took out loans similar to group loans.

Motivated in part by group lenders in less-developed countries, organizations in the United States have developed similar programs. The *1996 Directory of U.S. Microenterprise Programs* lists 51 organizations that issue group loans. The programs operate in both rural and urban areas. Often they are run by nonprofit organizations.

The underlying idea of group lending is to delegate monitoring and enforcement activities to borrowers themselves. Borrowers who know a lot about each other, such as those who live in close proximity or socialize in the same circles, are the most promising candidates for group lending. For example, the rural villages that Grameen lends in would seem ideally suited for group lending,

■ I would like to thank Hiroshi Fujiki, Tim Hannan, John Walter, Roy Webb, and John Weinberg for helpful comments. The views expressed in this paper do not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

because they are relatively self-contained communities, and people live close to each other and interact regularly. In such an environment, residents should be better than outsiders at assessing and monitoring the creditworthiness of fellow residents. They should also be better able to apply social pressure on potential defaulters.

The first goal of this paper is to analyze group lending, particularly as a potential method for lending to the poor in the United States. Studying alternatives to traditional lending is important because there is economic evidence that the poor in the United States have an unmet demand for finance. Zeldes (1989) finds that the poor are borrowing-constrained; that is, they would like to borrow more money at existing rates than they can. Evans and Jovanovic (1989), even after accounting for possible correlation between entrepreneurial ability and wealth, find that the lack of wealth affected the poor's ability to engage in self-employment activities. Bond and Townsend (1996), reporting on the results of a survey of financial activity in a low-income, primarily Mexican neighborhood in Chicago, find that bank loans are not an important source of finance for business start-ups. In their sample, only 11.5 percent of business owners financed their start-up with a bank loan. Furthermore, 50 percent of the respondents financed their start-up entirely out of their own funds.

Two services provided by financial intermediaries are delegated monitoring and asset transformation. Banks provide both of these services and, maybe surprisingly, groups do too. Group members monitor each other and through joint liability, transform the state-contingent returns of its members' loans into a security with a different state-contingent payoff. Consequently, groups can be interpreted as financial intermediaries, albeit small ones.

Interpreting groups as financial intermediaries is an important part of my second goal: to place group lending in the context of the rest of the financial intermediation sector. In this paper, groups have a comparative advantage at some types of financial intermediation. Understanding comparative advantage and specialization in financial intermediation to the poor is important because it can help answer questions such as: Which financial intermediary is best at what activity? How are different intermediaries financially linked? Do legal and regulatory restrictions, through their effect on the organization of financial intermediation sector, change the services they offer? These are the questions that underly the assessment of legislative acts aimed towards lending to the poor, like the Community Development Financial Institutions Act (CDFIA) and the Community Reinvestment Act (CRA).¹

¹ The two acts take different views on the importance of the structure of the financial intermediation sector for lending to the poor. The recently enacted CDFIA seems to take the view that alternatives to traditional financial institutions are needed to provide financial services to low-income communities. (See Townsend [1994] for a critical discussion of the act.) It funds institutions that specialize in providing financial services to low-income communities. In contrast,

Theoretical Framework

The theoretical framework I use is the delegated monitoring model developed in Diamond (1984, 1996). In his work, there are lots of small lenders and a smaller number of borrowers. Lenders lend to borrowers through a financial intermediary in order to economize on monitoring costs.

My model makes two additions to this framework; the major one is to allow some borrowers to monitor each other at a lower cost than outsiders. The heterogeneity in monitoring costs drives the coexistence of two types of financial intermediaries, large ones like banks and smaller ones like groups. Both types transform assets and provide monitoring services. In my model, just like in Diamond's model, lenders' funds flow through a large financial intermediary. But in my model, the large financial intermediary does not directly lend to all borrowers. Instead, for those borrowers who can monitor each other at a low cost, it lends to groups that in turn lend to their members.

I use Diamond's framework for three reasons. First, delegated monitoring is an important feature of group lending. Second, it allows for the embedding of groups into the financial intermediation sector. Finally, it demonstrates the similarities between groups and other financial intermediaries.

There is a small economic literature on group lending. This literature examines group versus individual lending but not in a model designed to study the existence of financial intermediaries. Stiglitz (1990) examines a problem where group members can assess whether other members are shirking. Varian (1990) examines the important screening role groups may provide, that is, their use of their prior knowledge about others to form groups. He also examines learning from fellow members as a potential advantage of groups. Besley and Coate (1995) examine the potential enforcement advantages groups may have. For example, social ostracism of defaulters is an option available to groups but not to outsiders. These penalties can reduce incentives to default but not in all cases. Sometimes, they increase the chance of default. While all of these features of group lending are important, I abstract from them.

In the following section, I provide background on group lending in practice, after which my model is developed and analyzed. Then I analyze the portability of group lending to the United States in the context of the model.

the CRA seems based on the premise that lending to low-income people is best done by existing financial institutions but that these institutions underserve low-income communities because of neglect, or even discrimination in the most egregious cases. The CRA works by requiring regulators to evaluate banks on criteria such as financial services provided to low-income communities. Banks that score poorly are subject to sanctions such as limits on merger activities.

1. GROUP LENDING IN PRACTICE

Microfinance is the provision of financial services to the poor. The prefix micro is used because the amounts involved in transactions are small. Often, microfinance is provided by nonprofit organizations; their targets are people who have not participated in the formal financial sector. The financial services that their clients *do* use tend to be supplied by relatives, or in some parts of the world, by moneylenders. Formal financial institutions have avoided this market because the loan sizes are small, administrative costs per dollar lent are high, and they perceive the risk of default to be significant. It is the absence of the formal sector from these markets that has led nonprofit organizations, often with the goal of poverty alleviation instead of profit maximization, to supply financial services. It is also the inappropriateness of traditional financial products that has led to the introduction of financial products such as group lending.

Group lending is not the only tool used to provide microfinance. Many microfinance organizations make loans only to individuals while others make loans to both individuals and groups. Others provide savings and insurance services. Much microfinance is provided informally, by rotating savings and credit associations, or between friends and family. While these issues are important, I do not discuss them because this paper is a study of the narrower question of what conditions favor group lending.

Group Lending in Less-Developed Countries

The most famous group lender is the Grameen Bank, which was founded in Bangladesh in the mid-1970s. This bank makes loans to groups of five unrelated individuals who are poor. Most groups consist of landless women from the same village. Loans are made sequentially with remaining members not receiving their loans until other members repay their loans. Loan size is increased after the group has successfully repaid earlier small loans.²

The bank has grown tremendously. In 1992, it lent to 2 million people at real interest rates of around 12 to 16 percent. Their repayment rate is high, around 97 to 98 percent. The bank even shows a profit, though it would not do so without the low-interest loans and grants it has received (Morduch 1997).

The Grameen Bank is far from the only institution to make group loans. Even in Bangladesh, there are at least two other organizations, Bangladesh Rural Advancement Committee (BRAC) and Thana Resource Development and Employment Programme (TRDEP), that make group loans (Montgomery,

² There are several other interesting features of the bank's organization. For example, collections of six groups are formed into Centres. All payments are made at Centre meetings in public view of other Centre members. Savings funds are also developed to provide for contingencies like death or disability. See Rashid and Townsend (1993) and Fuglesang and Chandler (1987) for more details.

Bhattacharya, and Hulme 1996). Like Grameen, BRAC is a sizable institution, lending to over 600,000 borrowers in 1992. Other countries with lending institutions that make group loans include Kenya (Mutua 1994), Malawi (Buckley 1996), Costa Rica (Wenner 1995), Columbia, and Peru, just to name a few.

One of the most successful group lenders is BancoSol, located in Bolivia. It is a chartered bank, subject to the supervision of SIB, the Bolivian bank regulatory agency. It makes uncollateralized loans for periods of 12 to 24 weeks. Repayments are made frequently, every week or two. Loans are made to what they call solidarity groups, each of which can have four to ten members. The group takes a loan from the bank and apportions it among its members. Like Grameen's groups, group members are jointly liable for each other's debts. Loans are usually made to provide working capital for small-scale commercial activities. Also like Grameen, the majority (77 percent) of clients are women. But unlike the Bangladesh bank, most of the borrowers are located in urban areas. Nonetheless, borrowers still have good information about each other because BancoSol requires all members of a solidarity group to work within a few blocks of each other. Most borrowers are market vendors, though half of the portfolio is lent to small-scale producers like shoemakers, bakers, and tailors (Glosser 1994). Lending is not the only financial service provided by BancoSol. It also offers deposit services in both boliviano and U.S. dollar-denominated accounts.³

BancoSol's growth has been extraordinary. In 1996, it lent to about 75,000 people, roughly one-third of the people who use the Bolivian banking sector. In 1996, BancoSol had a loan portfolio of \$47.5 million. It also earned \$1.1 million on revenues of \$13 million (Friedland 1997). Two important reasons for this success is that the bank charges real interest rates of 34 percent and has a default rate of less than 1 percent (Agafonoff 1994). The high interest rates are no doubt required to cover the high administrative costs required by its lending strategy. As a basis of comparison, 80 percent of BancoSol's costs are administrative, while the comparable number for the rest of the Bolivian banking industry is only 20 percent (Glosser 1994).

³ BancoSol is a chartered bank because Bolivian law requires deposit-taking institutions to be chartered banks subject to governmental supervision. BancoSol was created by PRODEM, a nonprofit organization that specialized in making loans. Its operations were financed mainly by grants, usually from foundations and USAID. The organization felt that grants were an insufficient source of capital, so it decided to create a regulated bank in order to have the legal right to collect deposits. Interestingly, the bank's nontraditional activities complicated the granting of the charter. For example, existing Bolivian banking law required that uncollateralized credit be less than twice paid-up capital. Unfortunately, for BancoSol, uncollateralized credit is all they supply! The bank negotiated a compromise in which loans under \$2,000 do not count towards this total. The costs of the conversion were not trivial. They exceeded \$500,000, according to one estimate (Glosser 1994).

Group Lending in the United States

Recently, several lenders have tried group lending in the United States. These lenders are nonprofit organizations whose main goal is to assist the poor—in particular, women and minorities—by financing self-employment. Since these efforts have started relatively recently, published information is still limited.

One source of information is a study by Edgcomb, Klein, and Clark (1996), who examine seven microenterprise programs. Of the seven, four make group loans.⁴ Each program provides services other than group lending. Several lend to individuals, others provide training, and some provide all three services.

All four programs followed Grameen's example but with modifications. Each agency started with groups of five members. However, the agencies found that if an individual dropped out of the group, the rest of the group would disband. Currently, three of the agencies allow more flexibility in group size. One program allows four to ten members, while another allows four to six businesses per group.

The scale of the agencies' operations are still small. For example, the number of loans made by the programs in 1994 ranged from 27 to 103, and average loan sizes ranged from about \$2,100 to \$4,900. Making these loans is expensive. The average cost per loan varied from \$4,500 to \$15,300, so these programs are far from self-sufficient. However, when compared with job training and other assistance programs, their costs seem more reasonable. I discuss possible reasons for the high costs after I describe the model.

Historical Group Lending

Group lending is often considered a recent innovation, and its recent popularity certainly is connected with the success of the Grameen Bank. There are, however, at least two types of institutions that existed long before the Grameen Bank and that used variants on group lending.

To the best of the author's knowledge the earliest institutions that used a form of group lending were the Irish Loan Funds (Hollis and Sweetman 1997a, b).⁵ The funds developed in the early 1700s, peaked in size in the early 1800s, and then slowly declined throughout the rest of the nineteenth century. Interestingly, Hollis and Sweetman trace their development to Jonathan Swift, the Anglican priest best known for writing *Gulliver's Travels*.

⁴ The four that made group loans were the Coalition for Women's Economic Development (CWED), based in Los Angeles; the Good Faith Fund (GFF), located in Pine Bluff, Arkansas; the Rural Economic Development Center (REDC), which lends throughout North Carolina; and the Women's Self-Employment Project (WSEP), based in Chicago.

⁵ All reported information about the Irish Loan Funds is taken from Hollis and Sweetman (1997a, b).

The Irish Loan Funds were usually located in rural areas, took deposits, and made small loans. The institutions generally made uncollateralized loans to finance a small investment project, such as the purchase of an animal. As a rule, the loans were repaid on a weekly basis. These loans most resembled present-day group loans in that all borrowers were required to obtain two cosigners for each loan, and both cosigners were liable for repayment.⁶ While each fund was independent, the funds were regulated by a Central Loan Fund Board.

Another historical example of European group lenders was that of the German credit cooperatives that developed in the late nineteenth century (Guinnane 1993; Banerjee, Besley, and Guinnane 1994). They were often located in rural areas where individuals knew each other well. These cooperatives provided credit services, and importantly, many had a policy of unlimited liability. That is, if the cooperative failed, any member could be sued for the entire amount owed by the cooperative. Interestingly, these credit cooperatives were the inspiration for the credit union movement in the United States.

2. THE MODEL

The model in this paper is designed to study the following three features of group lending and the financial intermediation sector:

- the existence of joint liability groups
- the existence of more traditional financial intermediaries
- large financial intermediaries lending to the groups

Analysis of these issues requires a model in which it is possible to lend funds either directly to an individual or indirectly through a financial intermediary. With two additions, the framework in Diamond (1984, 1996) provides an environment that satisfies these conditions.

Diamond considered an economy where there are borrowers and lenders, and funding each borrower's project requires the resources of several lenders. Borrowers' returns are unobserved by a lender unless he spends resources to monitor the borrower. Lenders face the choice of whether to lend directly to borrowers or to lend to them indirectly through the financial intermediary. In equilibrium, lenders lend to the financial intermediary and the intermediary in turn lends to the borrowers. The reason that lenders lend through the financial intermediary is that it avoids costly duplicative monitoring.

This paper operates in the same framework but with two additions, heterogeneous monitoring costs and screening costs. The important addition is the former. In particular, some borrowers are given the ability to form small groups,

⁶ The loans are much like the ones Swift made. Using his own money, he made small uncollateralized loans, required cosigners on loans, and required frequent repayments.

and in these groups they can monitor their fellow members. This ability is potentially valuable because group members monitor each other at a lower cost than a more traditional financial intermediary. People who live close to each other, those who work near each other, or those who socialize together would be most likely to satisfy these conditions. As in Diamond's model, it is optimal for lenders to lend to a traditional financial intermediary, but in this paper the financial intermediary lends to groups that in turn lend to their members. As we will see, the incentive problem underlying the contract between lenders and the large financial intermediary is the same as the incentive problem underlying the contract between the large financial intermediary and the groups. It is in this sense that groups and institutions, such as banks, are financial intermediaries for the same reason.

Environment

The model in this section is really a numerical example that closely follows Diamond (1996). In this economy, there are two main types of people, lenders and borrowers. Both types are risk-neutral, and consumption cannot go below zero. Each lender is endowed with $1/m$, $m > 1$, units of the investment good. The investment good cannot be consumed, but it can be used to create the consumption good. Lenders have access to a safe but low-return investment technology. Their investment technology takes x units of the investment good and turns it into $1.05x$ units of the good, receiving an interest rate of 5 percent.

The borrowers are better at producing the consumption good, but they start without any units of the investment good. Each borrower's investment technology requires an input of exactly 1.0 unit of the investment good. An investment of less than 1.0 produces an output of zero and any investment over 1.0 unit is wasted. The former assumption means that for each borrower it takes the funds of at least m lenders to finance his investment. Their investment technology is also riskier than that of lenders. In this example, an investment of 1.0 unit produces an output of 1.0 with a probability of 0.2 and an output of 1.4 with a probability of 0.8. Expected output for a borrower is $(0.2)1.0 + (0.8)1.4 = 1.32$, which is greater than 1.05, the return on the safe investment. However, 20 percent of the time output is less than what it would have been if the lender's investment technology had been used. Finally, I assume that each borrower's return is independent of other borrowers' returns.

In this model, the owners and the productive users of the investment good are different people. As the problem presently stands, the initial mismatch between owners and users is easily rectified through simple loan contracts. Lenders would lend to borrowers as long as their expected repayment was equal to 1.05. There is no role for intermediaries.

To introduce intermediaries requires the addition of complications to writing and enforcing contracts, complications that intermediaries are better able

to overcome than lenders. I now describe four features to the model that affect the feasibility and desirability of various contracts and ultimately lead to a role for financial intermediaries, both large ones like banks and smaller ones like groups. The four features are private information on borrowers' returns, liquidation costs, costly monitoring, and costly screening.

Private Information

It is assumed that borrowers' returns are private information. That is, a borrower is the only person who knows the success of his project; lenders do not observe it, nor do other borrowers. Private information makes some contracts infeasible. For example, consider a contract where lenders receive 1.0 if the low output is produced and 1.0625 if the high output is produced. If lenders knew that the contractual terms would be honored by the borrower, they would make the loan because their expected return is $0.2(1.0) + 0.8(1.0625) = 1.05$. Under private information, however, they cannot be sure that this contract would be honored. The reason is that lenders do not know the true value of the output so the borrower could always claim that he received a low output. That is, if the lender received the high output the borrower could claim he received the low output, pay 1.0 to the lender, and keep the difference. Lenders would be powerless to stop this deception; they cannot find out if he is telling the truth, and as things are presently specified, they cannot punish him. All they can do is refuse to lend, despite the acknowledged quality of his project.

Liquidation Costs

A contract with the option of liquidation is one way out of this dilemma. In this model, a liquidation cost serves as an ex post penalty imposed by the lender on the borrower. If the borrower does not meet the terms of his agreement, the lender can liquidate the borrower's assets. In this model, I interpret liquidating as meaning that the borrower and the lender receive zero. This means, among other things, that there are no assets that the lender can seize and sell. (In microfinance, projects are so small that one would gain very little from seizing and selling physical assets.)

The penalty imposed on the borrower by liquidation is important because it prevents him from always claiming he received the low output, as in the contract described above. For example, consider a debt contract with a face value F of 1.3125. If the borrower does not repay 1.3125, he has defaulted. When the output is 1.4, the borrower pays 1.3125. When the output is 1.0, the borrower cannot pay the full amount, so the lender liquidates, giving the borrower (and the lender) zero. The expected return to the lender is $(0.8)(1.3125) = 1.05$, so the loan is made and the borrower receives zero in the low-return state and 0.0875 in the high-return state. The threat of liquidation is enough to force repayment in the high-return state. The cost of liquidation is that output, which

is 0.2 in expected value terms, is destroyed. But in this example, the benefits of financing the loan outweigh the liquidation costs.⁷

Costly Monitoring

Costly monitoring is the other way to make lending feasible. In this paper there are two types of monitoring: costly monitoring by a lender and mutual monitoring within a group. Monitoring by a lender is identical to monitoring in Diamond's model; the lender pays an ex ante cost that allows him to observe a borrower's output. In essence, the lender uses resources to observe the private information. The resource costs could be as simple as spending time with the borrower or as complex as receiving regular reports on the project's financial status.

Observing output is valuable because then repayment can be made dependent on output, which avoids the need for liquidation. For example, consider the following contract: the lender monitors and the borrower pays 1.0 if the low output is realized, and 1.2 if the high output is realized. The expected return for the lender is $0.2 + 0.96 - K$, where K is the cost of monitoring. If the cost of monitoring is $K \leq 0.11$, then a lender's expected return (assume for the moment there is only one lender) is greater than 1.05, making monitoring worthwhile. Furthermore, this contract with monitoring is better for the borrower than the liquidation contract. (In both cases the borrower keeps zero in the low state, but under the monitoring contract, he keeps more in the high state.)

The second type of monitoring, mutual monitoring within a group, is the main departure from Diamond's model. I assume that within a subset of borrowers there are pairs of borrowers who know each other well, maybe because they live near each other or maybe because they are in the same social or ethnic circles.⁸ Each one of these pairs may form a group at a per-person cost of K_g . Membership in a group allows a group member to observe the other group member's output. Furthermore, because of the close social ties within a group, or maybe even because their time is less valuable than a loan officer's, I assume that the cost of being in a group is lower than the cost of anyone else monitoring them, that is, $K_g < K$.

At this point I should say more about what it means to be a group and how that affects the group's interaction with nongroup members. I am assuming that group members observe each other's outputs and act cooperatively or collude. In many models where people can collude, their interaction is complicated

⁷ There is no advantage from a contract that liquidates for the high output but not for the low output. Under such a contract, the borrower would always claim the low output, avoiding liquidation and keeping the difference between the high-output and the low-output payment. More generally, if the technology allows for more than two realizations of the output, even a continuum, then the optimal contract will still be a debt contract. The optimal contract will require a constant payment and liquidation if that payment is not made.

⁸ For simplicity, I assume that groups consist of only two people.

and even disadvantageous.⁹ In this model, there are no such disadvantageous effects. Furthermore, the analysis is simple because the borrowers are risk-neutral and thus utility is transferable. In this model, transferable utility eases the analysis because it means that the division of output between the group members does not affect the group's decisions. That is, regardless of how the group shares their returns, the group acts as if it is maximizing total expected output. In this paper, I assume that they share the returns equally. Besley and Coate (1995) examine a group-lending arrangement where there is an element of strategic play between group members, and they show that this can be a problem in some cases. I abstract from this consideration.

Screening Costs

The last element, and the remaining addition to Diamond's setup, is the addition of a screening cost. What I have in mind is a preliminary form of monitoring. A lender needs to meet with the borrower, discuss his project, and record and verify information about the borrower. In contrast with the previously discussed monitoring costs, screening costs do not reveal the final output. They only represent the effort that goes into ensuring that the project has a chance of success. To model these ideas, I assume that there is a fixed cost of K_s per lender to screen a borrower. I do not model what happens if the lender or lenders do not screen a borrower; I simply assume that they must screen a borrower before they make a loan.

I also assume that screening is only necessary for lending to borrowers. By borrowers I mean the second type of people, those who have access to the high return and risky technology, and not any entity that receives funds for investment. In particular, there is no need to screen a financial intermediary, though the financial intermediary still needs to screen any borrowers to whom it lends. This assumption is admittedly strong but not without merit. It seems reasonable to assume that it is harder to do a preliminary evaluation on small, idiosyncratic investment projects than on a large, well-known institution such as a bank. The only role of this assumption is to ensure that lending to groups is done by the financial intermediary and not directly by lenders.¹⁰

Where I am going . . .

In this economy there are lenders who have funds and borrowers who do not. The productivity of borrowers' investment projects creates a demand for

⁹ See, for example, Holmström and Milgrom (1990), Itoh (1993), Ramakrishnan and Thakor (1991), or Prescott and Townsend (1996).

¹⁰ There are other ways to ensure that lending to groups goes through the large financial intermediary, though they add additional issues that complicate the analysis. For example, making lenders risk-averse would be sufficient, since then each lender would want to lend directly to more than one group. Consequently, each lender would screen several groups, raising screening costs.

finance. Private information, however, precludes lending unless there is monitoring or the penalty of liquidation. Before describing how these elements create a demand for financial intermediation, it is helpful to show what the lending flows will be and where each type of financial intermediary fits into the flow pattern.

Figure 1 describes the direction of lending flows in the model. Arrows indicate the direction of lending and an M indicates whether or not there is monitoring. The lenders, who start with the investment good, make unmonitored loans to the large financial intermediary.¹¹ This financial intermediary makes two types of loans, monitored loans to individuals and unmonitored loans to groups. Groups, the smaller financial intermediary, in turn make monitored loans to its members.

My strategy for analyzing the model is to split the analysis into two sections. In the first section, I take as given that there is one large financial intermediary and analyze its decision of whether to make a loan to an individual or to a group. To do this analysis, I consider each type of loan the financial intermediary may make to the borrowers and enumerate the trade-offs of lending to a group versus lending to individuals and also whether or not it is beneficial to monitor the loans. Next, I consider the lending decisions for lenders and show that it is indeed optimal for them to lend to borrowers through the financial intermediary rather than to lend to them directly.

Lending by the Financial Intermediary

The large financial intermediary has three options for lending funds:

- It can lend to borrowers, not monitor them, and use the threat of liquidation;
- It can lend to borrowers and monitor them; or
- It can lend to borrowers through groups.

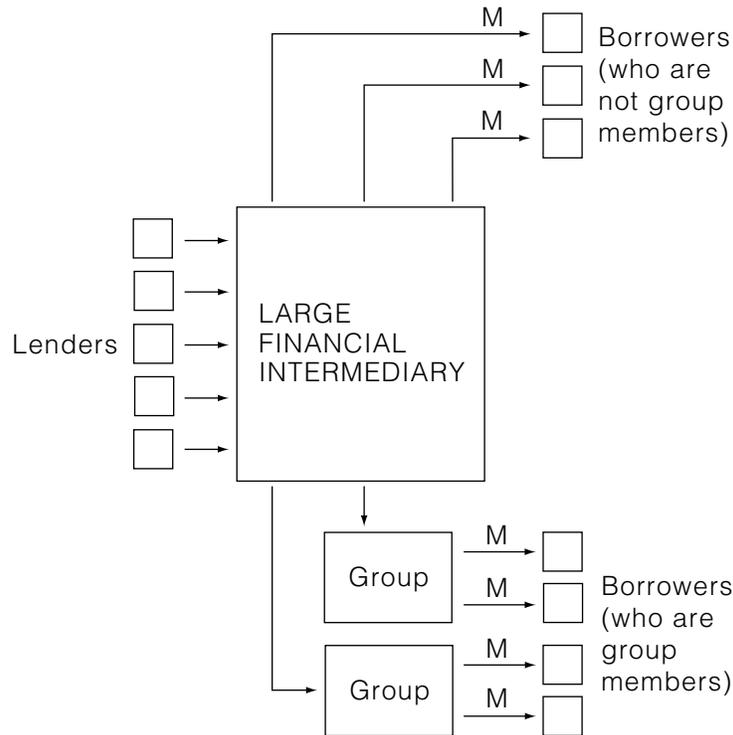
For this last case, we need only concern ourselves with unmonitored loans to the groups, since if the bank monitored them, it might as well bypass the groups altogether.

Recall that for each borrower who invests 1.0 unit of capital, he produces the low output of 1.0 with a probability of 0.2 and the high output of 1.4 with a probability of 0.8. Also, borrowers need 1.0 unit of the good to invest and for reasons explained later, the large intermediary requires an expected return of 1.05.

The expected returns to a project can be broken into five components: the expected payment to the financial intermediary R , the expected utility (return)

¹¹ There can be more than one large financial intermediary as long as each one has a sufficiently large portfolio. For our purposes, it is simplest to assume there is only one.

Figure 1 Lending Flows in the Model



Notes: M indicates that the loan is monitored. Arrows indicate direction of lending flows.

of the borrower U , the liquidation costs L , the monitoring costs M , and the screening costs S . These will sum to 1.32, the project's expected output. In the following sections, when each contract is analyzed, I will list the values of the five components for each contract. Also, I assume that the financial intermediary receives 1.05, the opportunity cost of the lenders' funds. Thus, any excess accrues to the borrower. Under this (unimportant) assumption, maximizing social welfare is equivalent to maximizing the utility to the borrower.

Individual Lending with Liquidation but No Monitoring

The enforcement device used for this contract is liquidation. Since there is no monitoring, a state-contingent contract without liquidation cannot be offered. Instead, a debt contract with a face value of F is written. Under this contract the borrower must pay F , or his project is liquidated. To make the problem interesting I assume that the parameters are such that $1.0 < F < 1.4$. This means that if the borrower receives the high return he pays F , but if he receives

the lower return, his project is liquidated, and both he and the intermediary receive zero. An F guaranteeing that the intermediary receives 1.05 in expected return is the solution to the following equation:

$$1.05 = (0.2)0 + (0.8)F - K_s.$$

The intermediary receives a zero payment 20 percent of the time, it receives a payment of F 80 percent of the time, and this return has to be high enough to cover the screening costs K_s and the opportunity cost of the funds 1.05. The solution to the equation is $F = 1.3125 + K_s/(0.8)$. The borrower's expected utility is $U = (0.2)0 + (0.8)(1.4 - F)$. Calculations for utility and the other variables of interest are as follows:

$$U = 0.07 - K_s,$$

$$R = 1.05,$$

$$L = 0.20,$$

$$M = 0, \text{ and}$$

$$S = K_s.$$

Notice that these values sum to 1.32, the project's expected return.

Individual Lending with Monitoring

There is no need to liquidate when monitoring because output is observed by the financial intermediary. For simplicity, I assume that 1.0 is paid out if the low return occurs, and F is paid out if the high return occurs. A face value of debt F that gives the intermediary a return of 1.05 is the solution to the following equation:

$$1.05 = (0.2)(1.0) + (0.8)F - K - K_s.$$

Compared with the previous contract, the intermediary now receives 1.0 if the low output is observed but must also bear the monitoring cost K . The solution to this equation is $F = 1.0625 + (K + K_s)/(0.8)$. The borrower's expected utility is again $U = (0.2)0 + (0.8)(1.4 - F)$. Carrying out the calculations for the variables produces the following numbers:

$$U = 0.27 - K - K_s,$$

$$R = 1.05,$$

$$L = 0,$$

$$M = K, \text{ and}$$

$$S = K_s.$$

Comparing the utilities from a loan with monitoring and a loan using liquidation shows that the former is preferred when $0.27 - K - K_s > 0.07 - K_s$, or equivalently, $K < 0.20$.

Group Lending

The group-lending contract includes elements of monitoring and liquidation. The group members monitor each other, but since the large financial intermediary does not know the results of their monitoring, it needs to include a liquidation provision in the contract. As I mentioned earlier, the two members of a group pool their resources so the group's distribution of returns is

return	probability
2.0	0.04
2.4	0.32
2.8	0.64

The assumptions made concerning group membership are that group members observe each other's output and act cooperatively. In this context, acting cooperatively means they maximize the expected value of the group's return. Thus, the contract needs to be written in terms of the total returns to the group, since the group can always move funds around to pay off a debt. Therefore, the optimal contract will again be a debt contract, with liquidation if the face value of the debt is not repaid. To facilitate comparison with the other contracts, we put the face value of the debt in per-group-member terms, that is, the face value of the group's debt is $2F$.

For the intermediary to receive an expected payment of 2.10 (1.05 per group member), F needs to solve the following equation:

$$2.10 = (0.04)0 + (0.32)(2F) + (0.64)(2F) - 2K_s.$$

I assume that the large intermediary rather than the group bears the screening cost. This assumption is not important.

At this point, it is necessary to make one more assumption. I assume that 2.4 units of output is enough to pay off the face value of the group's debt, $2F$. The value of $2F$ will depend on the other parameters, so I am assuming their values are such that this condition holds. Under these assumptions, the solution to the equation is $F = 1.09375 + K_s/(0.96)$. Each borrower's utility, assuming equal division of returns, is calculated from $U = (0.04)0 + (0.32)(2.4 - 2F - 2K_g)/2 + (0.64)(2.8 - 2F - 2K_g)/2$. I include the monitoring cost in this equation because the group pays it themselves. The values of the variables in per-group-member terms are

$$U = 0.23 - (0.96)K_g - K_s,$$

$$R = 1.05,$$

$$L = 0.04,$$

$$M = (0.96)K_g, \text{ and}$$

$$S = K_s.$$

Special attention should be paid to the liquidation cost, L . Under group lending, $L = 0.04$, which is dramatically lower than the case where the intermediary lends but does not monitor. (Recall that the liquidation cost in that case was 0.20.) The reason for the dramatic reduction is that the distribution of the group's output is different from the distribution of the individual's output. In particular, the group's distribution has less variance. The decreased dispersion of group returns reduces the incentive problem caused by the private information. In turn, a weakened incentive problem means that liquidation is invoked less often than a liquidation contract between the intermediary and an individual.

The argument is easier to understand if we compare two borrowers borrowing F each as a group with the same borrowers borrowing F each as individuals under the unmonitored liquidation contract. Also, assume that $1.0 < F < 1.2$. When the funds are lent to the individuals, each borrower's project is liquidated 20 percent of the time. This means that 4 percent of the time both are liquidated, 32 percent of the time one is liquidated, and 64 percent of the time neither is liquidated. Now compare these liquidation probabilities with those of the group. Under the group contract, 4 percent of the time both are liquidated, but 96 percent of the time neither is liquidated. The reason is that if one borrower gets a bad return and the other gets a good return, then the latter bails out the former. The transfers between the group members, in effect, alter their distribution of returns. This change reduces the probability of liquidation, which is beneficial.

One more way to view this problem, and an argument I will return to when discussing the large intermediary, is to consider a group consisting of a very large number of borrowers. (More formally, assume there is a continuum of them.) Because there are so many group members, the law of large numbers means that the group's total return is $1.32 - K_g$ with probability 1.0. All idiosyncratic risk averages out. In this case, there is never a need to liquidate since any claim that total output was less than $1.32 - K_g$ would not be credible.

To reiterate, groups greatly reduce the probability of being liquidated. Still, they have to pay a monitoring cost, and the relative size of these two costs (along with the intermediary's monitoring cost) determine whether group lending is better than the other types of lending. In this example, group monitoring is better than individual lending with monitoring if $0.23 - (0.96)K_g - K_s > 0.27 - K - K_s$; that is, the utility accruing to a borrower from group monitoring is greater than the utility accruing to a borrower from an individual lending with monitoring contract. Rearranging terms, the condition is

$$(0.96)K_g + 0.04 < K. \quad (1)$$

Equation (1) says that group monitoring is better if the sum of the group monitoring cost K_g and the liquidation cost of 0.04 is less than the intermediaries

monitoring cost K . This is not strictly true because K_g is multiplied by 0.96. That number, however, is only in the equation because groups bear the cost of monitoring; if their projects are liquidated, they receive zero and do not have to bear the monitoring cost.

I can now provide conditions under which the large financial intermediary will lend according to the pattern described by Figure 1. First, I assume that monitoring by the intermediary satisfies $K < 0.20$ (so individual lending with monitoring is better than individual lending without monitoring). Second, I assume that for some pairs of borrowers K_g is small enough to satisfy equation (1) and for other pairs of borrowers it is not. The former borrowers could be those who live near each other like Grameen's clients or work near each other like BancoSol's clients. For parameter values satisfying these conditions, borrowers who cannot form a group borrow as individuals with a monitored loan, while other borrowers who can form a group do so and borrow from the intermediary as a group, using the liquidation contract.

Lending to the Large Financial Intermediary

Now return to the lenders' lending decision. In equilibrium, as indicated by Figure 1, lenders lend to the large financial intermediary rather than directly to individuals or groups. Most of the pieces are already in place to demonstrate why this is the case. Lenders can either transform the asset themselves by using the low return but riskless technology, or they can choose one of the following four lending options:

- Lend directly to borrowers and use a liquidation contract;
- Lend directly to borrowers and monitor them;
- Lend directly to the group and use a liquidation contract; or
- Lend to the large financial intermediary.

The last option, lending to the large financial intermediary, is the optimal arrangement. I will demonstrate this by first showing that the costs to lenders of lending directly is greater than the same costs faced by the large financial intermediary making the same loans. Then, I will show that the lenders can lend to the large financial intermediary at no cost. This will mean that lending through the large financial intermediary is better than direct lending. Finally, if the large intermediary receives a return of 1.05, as was assumed in the previous analysis, and the intermediary adds no costs to lending, then it is optimal for lenders to lend to the large intermediary.¹²

The first three cases listed above are the direct-lending options available to lenders. Each one of these options corresponds to one of the cases worked

¹² Technically, lenders are indifferent between this option and using the safe investment technology. Among these two choices, I assume that the lenders choose the socially optimal one, which is to lend to the large financial intermediary.

through earlier in the section. The difference is that now monitoring and screening costs have to be borne by $m > 1$ lenders rather than just the large financial intermediary. The algebra is easy enough to work through but it is simpler to use the following observations. The incentives faced by a borrower do not depend on whether his funds are obtained from lenders or via the large financial intermediary. Consequently, the problem is unchanged from the earlier analysis except that screening costs (in all three cases) and monitoring costs (in the second case) are $m > 1$ times as much under direct lending. Therefore, it is cheaper for lenders to lend through the large financial intermediary rather than directly.

However, there still remains the issue of whether or not lenders need to monitor and screen the financial intermediary. If they do not, they can lend to the intermediary, which in turn lends to borrowers (either directly or indirectly through groups). This flow of funds will economize on monitoring and screening costs relative to direct lending.

By assumption, there is no need to screen the intermediary. However, some work is needed to demonstrate that lenders do not need to monitor the large financial intermediary. How do lenders know that the intermediary actually monitors the borrowers? How do they know the return of the intermediary? (At this point, it is helpful to think of the large intermediary as a person, possibly a lender, who if he did not monitor would save himself monitoring costs.)

In the previous section's analysis of lending to the group, the increased size of the group made the liquidation contract more effective. The larger the group, the more effective a liquidation contract was. If the group consisted of a continuum of members, then there was no need to monitor because the group's return is certain.

The same logic applies to the problem facing the lenders lending to the intermediary. If the intermediary lends to a continuum of borrowers, then the intermediary's return is certain. Thus, the optimal contract between lenders and the large financial intermediary is an unmonitored debt contract of face value $F = 1.05$. As part of the debt contract, the lenders liquidate the intermediary's assets if it claims its return is less than 1.05. But in equilibrium, the intermediary's portfolio is so diversified that its assets are never liquidated. Thus, there is no liquidation cost to lending through the financial intermediary, and there is no need to monitor it. The entire return of 1.05 that the intermediary receives from borrowers can be passed to the lenders. Lending through the large financial intermediary is better than direct lending.

To summarize, the large financial intermediary economizes on monitoring and screening costs while the groups economize only on monitoring costs. Relative to direct lending, both types of intermediaries economize on monitoring costs in the same way. Lending through the intermediaries avoids the duplicative monitoring of borrowers by lenders while the intermediary's diversification

reduces the need for lenders to monitor it. Thus, total monitoring is lowered in the economy. The reduction of these costs is the financial intermediary's special role in transforming assets.

There is, however, one way in which the two types of intermediaries differ in how they economize on monitoring costs. Compared with monitoring by the large financial intermediary, the groups save on monitoring costs because they have a cost advantage. It is efficient for the large financial intermediary to lend through groups if this cost saving outweighs the liquidation cost from using the group. The remaining observation—that lenders lend to groups through the large financial intermediary—occurs to economize on screening costs.¹³

3. ANALYSIS

Ideally, the model would be used in the following way. We would start with measurements of parameters in the model, such as distribution of returns, costs of monitoring, etc. These measurements would come from economies, like villages in Bangladesh or urban areas in Bolivia, where group lending is successfully used. Using these measurements we would evaluate the model on the criterion of whether or not it predicts there will be groups. If it does predict groups, the experiment proceeds by solving the model using parameter values taken from low-income U.S. communities. Then, the model could be used to evaluate the potential of group lending in the United States.

Precise measurement of many of these values is beyond the scope of this paper. Indeed, measurement of a concept like monitoring is a research project in and of itself. Consequently, the following discussion is necessarily sketchy, guided by what little information is available. Still, it is valuable, and one can gain some broad ideas about the role group lending and financial structure may play in channeling credit to the poor. The discussion should be considered a starting point, particularly for researchers and practitioners who are looking for guidance as to what variables to measure.

Business Opportunities

The model analyzes the problem of financing investment projects. It takes as given that potentially profitable investment projects exist. The financing

¹³ One difference from the Diamond (1984, 1996) setup is worth mentioning. In his paper, financial intermediaries exist only to economize on monitoring costs. In this paper, the large financial intermediary economizes on monitoring costs, but it also economizes on screening costs. The latter costs, in fact, are sufficient in this model for the large intermediary to exist. In this paper, monitoring costs serve the role of obtaining a nontrivial trade-off between individual and group lending. They are necessary to generate the existence of the small financial intermediaries, that is, the groups.

problem, however, is irrelevant if there are no profitable microenterprise projects to finance.

The evidence presented in the introduction suggests that there are profitable investment projects in the United States that would be financed in the absence of information constraints. There are, however, reasons to think that there may be less of these opportunities in the United States than in Bangladesh or Bolivia. For example, in less-developed countries 60 to 80 percent of the labor force is engaged in self-employment (Edgcomb, Klein, and Clark 1996), while in the United States only about 12 percent of the labor force is self-employed (Segal 1995). Ultimately, of course, the existence of profitable self-employment opportunities must be determined by empirical investigation.

A related issue, applicable to most microfinance programs, is what type of investments can be financed with group lending or any other microfinance program. For example, one key feature of the studied lending programs is the required frequency of repayments. Frequent repayment requires that an investment produce cash flow for the entire course of the loan. If it does not, then the borrower will default. This time path would seem to preclude loans for investments that pay off sometime in the future. For example, a planting loan to a farmer is poorly suited for frequent repayment because planting does not generate income until harvest.

A cursory examination of the type of loans made by Grameen, BancoSol, or the Irish Loan Funds bears out this observation. Despite their rural location, planting loans are not made by Grameen nor were they frequently made by the Irish Loan Funds. Many loans tend to be for investments that produce a flow of income. The purchases of a cow that produces milk or a chicken that lays eggs are examples of such an investment. BancoSol's loans, while in a different context, serve a similar purpose. They tend to be made for working capital.

Conceivably, there are many valuable investments that do not produce the steady cash flow demanded by group and other microfinance lending schemes. The important question here is why are the loans made with these terms? Are frequent repayments an important part of monitoring? The answers to these questions are important not just to the evaluation of group lending in the United States but also for the evaluation of lending in less-developed countries.

Source of Funding and Comparative Advantage in Lending

The source of funding is important because it can limit the activities of a financial intermediary, and it can influence the optimal structure of the financial intermediation structure. In the model, there were many lenders per borrower. This ratio was responsible for the existence of the large financial intermediary since the number of lenders needed to finance a borrower determines the costs of direct lending, and consequently the savings in monitoring and screening

costs from intermediation. For microfinance programs it is reasonable to ask if there are lots of lenders per borrower. First, the loans are for small amounts, and second, many lenders are donors with large amounts to lend.

BancoSol receives some of its funding from deposits. Agafonoff (1994) reports that in 1994 BancoSol's average loan was \$499 and its average deposit was \$225. (The majority of the bank's loans and deposits are denominated in U.S. dollars rather than Bolivian bolivianos.) These numbers are consistent with the model's assumption.

Still, many investors are large organizations whose investments are much higher than the amount any single individual borrows. In terms of the model some modifications would need to be made to ensure that donors lend through an intermediary rather than directly. The simplest, and most obvious, would be to assume that donors do not have the expertise to lend themselves. Consequently, K and K_s are much higher if they lend themselves rather than through an intermediary. Another possibility is that donors, particularly those overseas, find it expensive to monitor because of physical, linguistic, and even cultural distance from the borrowers. (See Boyd and Smith [1992] for a model in which people at different locations have a comparative advantage in lending in their home location.)

A comparison of the United States and Bolivia suggests that a group lender may desire different sources of funds in the two countries. In Bolivia, BancoSol raises some of its funds from deposits, but it is a country where a large fraction of the population does not use the banking sector. The banking sector, and more generally the financial structure, is much more extensive in the United States. Consequently, raising deposits might not be a group lender's comparative advantage. Instead, debt or equity might be a better source of capital for a group lender in the United States.

In the United States, group lenders' comparative advantage should be in lending rather than in collecting deposits. Lending to the poor likely requires a different set of skills than other types of lending. BancoSol's high administrative costs relative to the rest of the Bolivian banking sector is supportive of the latter conjecture.¹⁴

Indeed, it is not difficult to imagine a highly specialized financial system where traditional financial intermediaries collect deposits and then direct funds to specialists in microfinance, who in turn lend to groups (or individuals). There is no reason to think that traditional financial intermediaries are the best institutional vehicles for delivering credit to the poor.

¹⁴ In the model, groups save on monitoring costs, yet in the data, group lenders spend a lot of resources on monitoring. This is not a contradiction. The issue is how much more resources would have to be used to monitor in the absence of groups. That is what the model captures.

Costs to the Large Intermediary

In the model, for some parameter values the large intermediary saved costs relative to direct lending. In practice, monitoring and screening costs may be so high as to make any form of financing unprofitable. The problem is particularly acute for microfinance because loans are for small amounts, and they require frequent repayments. In the context of the model's parameters, K and K_s might be much higher in the United States than in less-developed countries.

The data bears out the importance of these costs. Eighty percent of BancoSol's costs are administrative while the cost figures for the U.S. agencies exceed the face value of the loans. BancoSol has surmounted these problems through a combination of a low default rate and a high interest rate (about 34 percent per year). In 1994, their average cost per dollar lent was 0.16; their borrower-to-loan-officer ratio was about 320.

Any microfinance program in the United States that desires to even approach self-sufficiency will need a similar strategy and results. None of the four agencies have reached BancoSol's scale. No agency made more than 107 loans in 1994. Their loan-loss ratios vary from about 2 to 17 percent, and their costs per dollar lent uniformly exceed one. These programs are far from self-sufficient. Of course, these programs are relatively new and any activity takes time to learn, not to mention the time needed to obtain economies of scale. It would be interesting to compare these agencies' default rates with those of Grameen or BancoSol in their early years of operation.¹⁵

Still, self-sufficiency may be too strong an evaluative criterion. Many services and transfers are distributed through the social welfare system and these programs are the right basis for comparison. Under this interpretation, microfinance is unusual in that it directs aid to specific people in the population; those who are willing to start businesses. Furthermore, unlike most social welfare programs, the recipients face the explicit incentive to perform or lose their aid. Under this criterion, group lending may very well be an effective method for targeting aid to the poor, particularly since these agencies' costs are comparable with those of job-training programs.

Monitoring within Groups

One of the most critical issues concerning group lending is how high is K_g , the cost of group monitoring?¹⁶ There are reasons to think that K_g is higher in

¹⁵ A potential problem for any program with the goal of self-sufficiency is that the interest rates necessary to cover costs may be illegal, violating usury laws in many states of the union.

¹⁶ In the model, monitoring was an either-or proposition. The only options available were to pay the monitoring costs and observe fellow members' output or to not pay the cost and not see the output. In practice, there are degrees of monitoring. Still, for the purposes of our discussion, K_g provides a useful way to summarize these degrees.

the United States than in developing countries. There is more anonymity, the costs of being excluded from a group are smaller in a rich country, and people do not necessarily work in such close quarters.

Edgcomb, Klein, and Clark (1996) provide some indirect evidence in support of this view. They conclude that the group-lending programs have had the most trouble in rural areas. The programs found that rural residents do not tend to know each other well enough to be able to support groups, in part because of the low density of the population and in part because of the low number of self-employed people in rural areas. One agency has even resorted to purchasing credit reports on fellow members for potential groups.

Another complication is that self-employment opportunities are more diverse in the United States than in less-developed countries (Edgcomb, Klein, and Clark 1996).¹⁷ Group members engaged in similar activities can learn from each other and can evaluate the borrowing proposals of fellow group members. It probably also makes monitoring easier. This is another reason K_g may be higher in the United States. Some of the resources used on training by the U.S. programs may be designed to compensate for this.

4. CONCLUSION

Lending groups are financial intermediaries, albeit small ones. The model shows how groups, as well as larger financial intermediaries, economize on monitoring costs and transform assets. Through diversification, financial intermediaries alleviate incentive problems and reduce the costs of monitoring and screening.

Throughout the paper, I provide extensive description of existing group-lending programs to demonstrate that group lending is a type of intermediation that is viable in at least several environments, including some of older origin than many probably realized. Whether it is viable in the United States is an open question, though the conditions here appear to be less favorable for it than in less-developed countries. Still, while the narrow focus of this paper is on the relative merits of group lending, the broader goal is to study financial structure. Understanding financial structure is a necessary prerequisite to the proper formulation of policy involving financial intermediation and low-income communities.

¹⁷ However, different activities may have less-correlated returns. In my model, group lending is more valuable when returns are less correlated.

REFERENCES

- Agafonoff, Alexander. "Banco Solidario S.A.: Microenterprise Financing on a Commercial Scale in Bolivia," Economics Division Working Paper 94/5. Research School of Pacific and Asian Studies, The Australian National University, 1994.
- Banerjee, Abhijit V., Timothy Besley, and Timothy W. Guinnane. "Thy Neighbor's Keeper: The Design of a Credit Cooperative with Theory and a Test," *Quarterly Journal of Economics*, vol. 109 (May 1994), pp. 491–516.
- Berenbach, Shari, and Digo Guzman. "The Solidarity Group Experience Worldwide," in Maria Otero and Elisabeth Rhyne, eds., *The New World of Microenterprise Finance*. West Hartford, Conn.: Kumarian Press, 1994.
- Besley, Timothy. "Savings Credit and Insurance," in J. Behrman and T. N. Srinivasan, eds., *Handbook of Development Economics*, Vol. III. Amsterdam: Elsevier, 1995.
- _____, and Stephen Coate. "Group Lending, Repayment Incentives and Social Collateral," *Journal of Development Economics*, vol. 46 (February 1995), pp. 1–18.
- Bond, Philip, and Robert M. Townsend. "Formal and Informal Financing in a Chicago Ethnic Neighborhood," *Federal Reserve Bank of Chicago Economic Perspectives*, vol. 20 (July/August 1996), pp. 3–27.
- Boyd, John H., and Bruce D. Smith. "Intermediation and the Equilibrium Allocation of Investment Capital: Implications for Economic Development," *Journal of Monetary Economics*, vol. 30 (December 1992), pp. 409–32.
- Buckley, Graeme. "A Study of the Malawi Mudzi Fund and the Smallholder Agricultural Credit Administration," in David Hulme and Paul Mosley, eds., *Finance Against Poverty*, Vol II. London: Routledge, 1996.
- Diamond, Douglas W. "Financial Intermediation as Delegated Monitoring: A Simple Example," *Federal Reserve Bank of Richmond Economic Quarterly*, vol. 82 (Summer 1996), pp. 51–66.
- _____. "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies*, vol. 51 (July 1984), pp. 393–414.
- Edgcomb, Elaine, Joyce Klein, and Peggy Clark. *The Practice of Microenterprise in the U.S.: Strategies, Costs and Effectiveness*. Washington: The Aspen Institute, July 1996.
- Evans, David S., and Boyan Jovanovic. "An Estimated Model of Entrepreneurial Choice under Liquidity Constraints," *Journal of Political Economy*, vol. 97 (August 1989), pp. 808–27.

- Friedland, Jonathan. "Bolivian Bank Thrives with Little Loans," *Wall Street Journal*, July 15, 1997.
- Fuglesang, A., and D. Chandler. "Participation as Process: What We Can Learn from the Grameen Bank, Bangladesh." Norwegian Ministry of Development, 1987.
- Glosser, Amy J. "The Creation of BancoSol in Bolivia," in Maria Otero and Elisabeth Rhyne, eds., *The New World of Microenterprise Finance*. West Hartford, Conn.: Kumarian Press, 1994.
- Guinnane, Timothy W. "Cooperatives as Information Machines: The Lending Practices of German Agricultural Credit Cooperatives, 1883–1914," Economic Growth Center Discussion Paper No. 699. Yale University, August 1993.
- Hollis, Aidan, and Arthur Sweetman. "Microcredit in Pre-Famine Ireland." Economics Working Paper Archives, ewp-eh/9704002. April 1997a.
- _____. "Complementarity, Competition and Institutional Development: The Irish Loan Funds through Three Centuries." Economics Working Paper Archives, ewp-eh/9704003. March 1997b.
- Holmström, Bengt, and Paul Milgrom. "Regulating Trade among Agents," *Journal of Institutional and Theoretical Economics*, vol. 146 (March 1990), pp. 85–105.
- Itoh, Hideshi. "Coalitions, Incentives, and Risk Sharing," *Journal of Economic Theory*, vol. 60 (August 1993), pp. 410–27.
- Krasa, Stefan, and Anne P. Villamil. "A Theory of Optimal Bank Size," *Oxford Economic Papers*, vol. 44 (October 1992), pp. 725–49.
- Montgomery, Richard, Debapriya Bhattacharya, and David Hulme. "Credit for the Poor in Bangladesh," in David Hulme and Paul Mosley, eds., *Finance Against Poverty*, Vol. II. London: Routledge, 1996.
- Morduch, Jonathan. "The Microfinance Revolution." Manuscript. 1997.
- Mutua, Alvert Kimanthi. "The Juhudi Credit Scheme: From a Traditional Integrated Method to a Financial Systems Approach," in Maria Otero and Elisabeth Rhyne, eds., *The New World of Microenterprise Finance*. West Hartford, Conn.: Kumarian Press, 1994.
- 1996 Directory of U.S. Microenterprise Programs*. Self-Employment Learning Project. Washington: The Aspen Institute, 1997.
- Prescott, Edward Simpson, and Robert M. Townsend. "Theory of the Firm: Applied Mechanism Design," Federal Reserve Bank of Richmond Working Paper No. 96–2. June 1996.
- Ramakrishnan, R. T. S., and A. V. Thakor. "Cooperation versus Competition in Agency," *Journal of Law, Economics, & Organization*, vol. 7 (Fall 1991), pp. 248–83.

- Rashid, Mansoor, and Robert M. Townsend. "Targeting Credit and Insurance: Efficiency, Mechanism Design, and Program Evaluation." Manuscript. 1993.
- Segal, Lewis M. "Flexible Employment: Composition and Trends," Working Paper 95-19. Chicago: Federal Reserve Bank of Chicago, December 1995.
- Srinivasan, Aruna. "Intervention in Credit Markets and Development Lending," Federal Reserve Bank of Atlanta *Economic Review*, vol. 79 (May/June 1994), pp. 13-27.
- Stiglitz, Joseph E. "Peer Monitoring and Credit Markets," *World Bank Economic Review*, vol. 4 (September 1990), pp. 351-66.
- Townsend, Robert M. "Community Development Banking and Financial Institutions Act: A Critique with Recommendations," in *Proceedings of the 30th Annual Conference on Bank Structure and Competition* (Federal Reserve Bank of Chicago, May 11-13, 1994), pp. 538-46.
- Varian, Hal R. "Monitoring Agents with Other Agents," *Journal of Institutional and Theoretical Economics*, vol. 146 (March 1990), pp. 153-74.
- Wenner, Mark D. "Group Credit: A Means to Improve Information Transfer and Loan Repayment Performance," *Journal of Development Studies*, vol. 32 (December 1995), pp. 263-81.
- Zeldes, Stephen. "Consumption and Liquidity Constraints: An Empirical Investigation," *Journal of Political Economy*, vol. 97 (April 1989), pp. 305-46.

Investing in Equities: Can it Help Social Security?

Michael Dotsey

Social Security is in trouble. A recent report by the U.S. General Accounting Office (1997) indicates that absent any changes to the current system, payments to beneficiaries will exceed revenues from payroll taxes in 2012, and by 2029 the Social Security Trust Fund will be depleted. That Social Security is in trouble is not really news. The system has a long history of being underfinanced, and the current difficulties are not historically large. Recently, the 1994–1996 Advisory Council on Social Security issued its report with various recommendations for putting the system on firm financial footing. From an economic perspective, making the Social Security System sound is not a difficult task. There exist a multitude of ways for doing so, but most involve either increases in taxes, reductions in benefits, or both. Thus, any plan inherently involves difficult political decisions. However, one part of the solution that is included in each of the three separate plans that were presented to the Commissioner of Social Security was the recommendation that some portion of the current trust fund be invested in the stock market. By taking advantage of the higher returns earned by equities, this recommendation seemingly would reduce the increases in taxes or the reduction in benefits that would be needed to return the Social Security System to financial viability.

In this article I address the economic merits of this recommendation. My analysis suggests that the ownership of the capital stock has very few consequences for the government's budget. The economic opportunities available to society are not increased by a transfer of capital from the private sector to the government. In short, there is no free lunch.

■ I wish to thank Douglas Diamond, Andreas Hornstein, Thomas Humphrey, Kent Smetters, and Alex Wolman for many useful suggestions and comments. The views expressed herein are the author's and do not represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

1. A BRIEF HISTORY

The Inception of Social Security

Social Security was created in 1935 as an intergenerational transfer program from workers to retirees. Its design also provided for income redistribution among the elderly, because replacement rates (the ratio of the benefit paid in the first year of retirement to taxable earnings in the preceding year) are higher for low-income workers than for high-income workers. Social Security is a pay-as-you-go system. In years when the revenue from Social Security taxes exceeds outlays, the U.S. Treasury uses the proceeds to finance other expenditures, thereby reducing the level of government debt from what it otherwise would have been. There exists the accounting fiction of a trust fund, but from an economic perspective no such fund exists. The lower level of government debt makes it more likely that future claims will be honored, but there is no dedicated set of securities belonging to the Social Security Administration that has the same legal standing as a government bond issued to a private citizen. Because it is a pay-as-you-go system, there is the potential that, for a variety of reasons, promised payments could become increasingly difficult to honor. This is what has happened repeatedly to the Social Security System.

Despite its problems, the Social Security System has been remarkably successful in terms of its growth and its economic importance. At the time of its creation, the old age and survivors insurance (OASI) part of the program was fairly small, with benefits equaling 0.03 percent of GDP in 1940. By 1950 that percentage had risen to only 0.33 percent of GDP, but by 1996 it had risen to over 4 percent of GDP. In terms of taxable payrolls, benefits were 10.7 percent in 1994, which is very close to the 10 percent envisioned in the 1939 Act (see Miron and Weil [1997]) and represented roughly 19 percent of all federal outlays. Over time, the fraction of the labor force covered by Social Security has risen from 63.7 percent in 1940 to 97.6 percent in 1993.

Social Security has also played a major role in reducing the percent of those over 65 who live below the poverty line. In 1959, 35.2 percent of the elderly were characterized as poor. By 1994, that figure had dropped to 11.7 percent. The increase in Social Security benefits is in large part responsible for this decline. Expressed in terms of 1995 dollars, the average monthly benefit in 1950 was \$269.30, in 1960 it was \$381.38, and in 1995 it was \$719.80. Also, the number of beneficiaries has risen substantially from 222,488 in 1940 to 37.5 million in 1995. In terms of percentages of the population over 65, only 7 percent received benefits in 1940, whereas 91.3 percent received benefits in 1995. More importantly from the standpoint of helping the poor, Social Security currently provides over 90 percent of income for half the seniors below the poverty line and 50 percent of income for two-thirds of all beneficiaries.

As the preceding figures show, the scope and amount of coverage has increased greatly since the inception of Social Security. The original act

promised benefits only to those who contributed, but in 1939 benefits were extended to spouses and surviving widows. Over time, various changes have expanded the scope of Social Security, with perhaps the most important extension resulting from the 1950 Act that brought 10 million new workers into the system. Also, various changes in computing benefits, coupled with high inflation and growth in wages, served to increase benefits, which consequently grew much faster than the economy.

Initially a 2 percent tax rate, equally divided between employer and employee, was levied on income up to \$3,000. The first benefits to contributing retirees were not to be paid until 1942, but the 1939 Act moved that date forward to 1940. Further, no benefits were to be paid in any month that a retiree earned more than \$15. To put that figure in perspective, the average annual wage in 1937 was \$979. This feature of the system indicates that Social Security was in part envisioned as insurance against destitution. However, under the assumption of no inflation and no wage growth, the replacement rate for a worker earning \$1,000 for 45 years, and retiring at age 65 in 2002, would have been 0.60 under the initial act. That means that this hypothetical worker would have received \$600 a year in perpetuity, implying that the initial act also possessed features that went far beyond mere insurance. A 60-year-old worker earning the same salary (\$1,000) and retiring in 1942 would have received benefits of \$200 a year. With the extension of benefit eligibility, the 1939 Act also reduced the replacement rate to 0.43. Thus, our hypothetical worker would receive only \$430 upon retirement, while his spouse would receive \$215.

Under the 1939 Act, the combined tax rates on employer and employee were 2 percent and were scheduled to rise to 6 percent by 1949 and remain fixed thereafter. Full benefits would not begin until 1991, when workers with a full history of contributions would be retiring. According to projections at the time, the internal real rate of return for those retirees would be 3.9 percent, not much above the 3 percent rate of return that was projected on the accumulated trust fund. Or, in more relevant terms, the internal rate of return would not be too far above the economy's growth rate and benefits could be paid by issuing government debt without increasing the debt-to-GDP ratio. Thus the initial planning attempted to create a sustainable system.

A History of Problems

Over its history the Social Security System probably has never been sound. The chief reason is that Congress tended to make benefits more generous than originally intended and refused to raise tax rates as fast as the 1939 Act prescribed. Tax rates did not reach 6 percent until 1960. Also, economic factors such as usage growth interacted with the methodology for calculating benefits, increased the level of benefits in unintended ways during the 1970s, and placed the system under tremendous strain. Corrections to the methodology were not

made quickly enough, and tax rates were not raised sufficiently, so that the system almost defaulted in the early 1980s. Demographic changes also conspired to make the system less sound than it would have been under stable population growth. Thus, under current law someone just entering the labor force will earn a rate of return on Social Security contributions that is probably negative, while the rate of return for those that have already retired is significantly higher than was intended.

The 1950 Act, which brought in 10 million new workers, also calculated their benefits in a way that provided them with large transfers. Expansion in scope need not have been detrimental to the soundness of the system, but these workers received benefits that were based on their wage history after 1950 rather than on their entire wage history. Thus, individuals from this group who retired soon after 1950 received full benefits and a large transfer from the existing system. Basically for this group the link between the replacement rate and the number of years of paying into the system was cut, and these new retirees received the same benefits as those who had been in the system since its inception. To accommodate this change, average benefits were slightly reduced.

Perhaps the most severe problem for the system was created by the 1972 Act, which for the first time included automatic price adjustments. Previously, such adjustments were made on an ad hoc basis. However, the adjustment procedure ended up overcompensating workers and made replacement rates unstable (for an excellent discussion see Munnell [1977]). The cost-of-living adjustment for retirees did not present a problem. Rather, the calculated replacement rates for newly retired workers were overstated. In essence these workers received an increase in their benefits that accounted not only for inflation but for wage growth as well. Because wages tend to rise with inflation, new retirees received a double counting. The amount of initial benefits also increased with the disparity between real wage growth and inflation. In this manner, the economic climate at the time, along with the unsound method of computing initial benefits, placed great stress on the system, with replacement rates rising from 47.9 percent in 1970 to 66.7 percent in 1980. As a result, the individual that retired in the 1970s received the largest net transfer of any cohort under Social Security.

The mistakes in the 1972 Act led to the rescue package of 1977, which constituted the largest peace-time tax increase in U.S. history. The rescue package also stopped initial benefits from rising faster than wage growth. The system was pronounced sound for the rest of the century and well into the next one. Unfortunately, the pronouncement was wrong. By 1981, there was a high probability that the system would not be able to meet its promised benefits. A commission was appointed to deal with the problem. Its lack of complete success is in part why Social Security restructuring is currently receiving so much attention. The 1983 Act did raise the schedule of tax rates and the annual

maximum on taxable earnings. It also effectively reduced benefits by taxing some portion of Social Security payments. Finally, it gradually raised the age to 67 at which full benefits were paid for the cohort born in 1960. Combined, these changes averted a problem of failing to honor legislated benefits but failed to solve the problem of long-term insolvency.

The Current Problem

The Social Security System as currently constituted is not actuarially sound. In this regard, the important date is 2012, because that is when expenditures will exceed receipts. At that point the federal government will have to raise taxes, reduce government spending, or increase its borrowing in order to make the promised payments to retirees. Beyond that date, the revenue shortfall will increase and the necessary adjustments will be more dramatic. It is estimated that the revenue shortfall will be \$57 billion in 2015 and grow to \$232 billion in 2020. Put in perspective, current total OASI payments are approximately \$308 billion dollars. This deficit will occur in part because there will be an estimated 50.4 million beneficiaries in 2015, up from 37.5 million in 1995.

As mentioned earlier, the system's current troubles are a consequence of increasing benefits, due both to the increased number of retirees and the more generous benefits that each retiree receives. One way to gauge the increase in the level of benefits is to compare them with average wages. For example, in 1953 the maximum benefit was equivalent to 30.5 percent of the average wage. By 1981 the corresponding figure was greater than 50 percent, and in 1995 it equaled 60.5 percent (Marcks 1997). Unquestionably, retirees' benefits have been rising relative to the tax base that can support those benefits.

The problem is also one of demographics. In 1945 there were 42 workers per retiree. In 1995 that number had shrunk to 3.3, and it is projected that in 2030 there will only be 2 workers per retiree. Furthermore, the life expectancy of individuals has increased since the inception of the system, meaning that a greater fraction of contributors have become beneficiaries. Also, the length of retirement has increased. In 1940 a 65-year-old male and female had a life expectancy of 12 and 13 years, respectively. By 2015 the comparable numbers will be 16 and 20 years.

These demographic features imply that maintaining the current level of benefits requires a significant increase in taxes. The *Report of the 1994–1996 Advisory Council on Social Security* (Department of Health and Human Services 1997) indicates that taxes would have to be raised immediately by 2.13 percent to attain 75-year balance.¹ These calculations explicitly take into ac-

¹ It should be noted that 75-year balance and actuarial soundness are not the same thing, because the problems of the system tend to worsen in the future. Thus 75-year soundness today implies 75-year unsoundness tomorrow.

count interest payments and payments on principal from the fictitious trust fund. To make these payments, the government would have to increase the level of the debt, reduce spending, or increase tax revenue from other sources.² Thus, total tax payments could be substantially higher if all forms of taxes are considered. Waiting to adjust tax rates will only make the problem worse.

2. THE STOCK MARKET TO THE RESCUE?

Over the period 1926 to 1993, the real return on the Standard & Poor's 500 averaged 9 percent, while the real yield on intermediate-term U.S. government bonds averaged only 2.2 percent. This difference in yields is large. For example, earning 9 percent implies that your investment doubles approximately every eight years as compared to every 35 years with a 2.2 percent return. Furthermore, in every 22-year period since 1926, equities have outperformed bonds. These considerations have spurred many observers to argue that investing at least some portion of the Social Security Trust Fund in equities can avert the financial difficulties facing the system.

In one sense the proposition is true. Increasing the yield on the trust fund can make the Social Security System more viable in isolation. However, it can only do so by making the rest of the government worse off. On net, an individual taxpayer will be little affected by this investment policy. In order for the government, or some part of it, to take an equity position, the government as a whole must issue more bonds. This swap of paper claims with the public affects the allocations and the risk characteristics of the respective portfolios but quantitatively does not have any appreciable effect on the government's overall budget. The economy cannot produce any more goods, and although the consumption profile of the representative household may be somewhat altered, the effects of this alteration are small. Since taxpayers are the ultimate receivers of any government earnings or losses, it matters little who owns the capital stock.

A number of economists recognize this fairly simple notion. Federal Reserve Board Chairman Alan Greenspan expressed the idea cogently in his recent Remarks at the Abraham Lincoln Award Ceremony of the Union League of Philadelphia (1996, p. 8), "Bonds and equities are merely the paper claims to income earning assets, and the value of the income stream is not determined by short-run changes in the supply and demand for securities. Rather, equity prices

² If the payments promised by Social Security are equivalent to payments promised on government bonds, then increasing the level of the measured debt to pay off these claims does not affect the overall indebtedness of the U.S. government. This action just transfers a promise into an explicit security. Treating the promised Social Security benefits in a similar way to any other government IOU implies that the true level of the government debt is closer to \$17 trillion instead of the \$5 trillion currently calculated.

must, in the long run, reflect the underlying earnings of the corporations on which the equities are a claim, as well as society's need to be compensated for postponing consumption into the future and its perception and attitudes toward risk as a consequence of uncertainty about the future. Indeed, the total market value of debt plus equities is, to a first approximation, likely to be unaffected by a shift in the balance of paper claims." These sentiments are also reflected in the views of Herbert Stein (1997, p. A18), ". . . privatizing the Social Security funds would not add to national saving, private investment, or the national income and would not allow the system to earn more income without anyone earning less."

Others, however, have argued to the contrary and have made the purchase of equities by the trust fund seem like a free lunch. For example, editorial commentary in *Barron's Online* by Thomas G. Donlan (1997) states that "Unless the system invests in private enterprise and those investments continue to earn historically high returns the Baby Boom generation will pay for its own retirement." Investing in equities is a major component of all three plans presented by the 1994–1996 Advisory Council on Social Security (Department of Health and Human Services 1997).

The Trust Fund

The Social Security System is but one part of the government. It is the largest part, with transfers amounting to 22 percent of government expenditures in 1995. The system's trust fund is really a myth. Social Security receives contributions or taxes from workers and their employers and pays out benefits to retirees, their dependents, and those on disability. Excesses in receipts over expenditures are handed over to the U.S. Treasury to be used in financing other governmental activities. Employing an accounting fiction, the Social Security System treats these transfers as investment in government securities and adds them to an imaginary portfolio that also collects fictitious interest payments. From the perspective of the government's total budget, this practice implies that the Treasury issues fewer bonds to the public than it would if there were no surplus received from Social Security. Unlike Treasury bills issued to the public, however, the IOUs from the Treasury to Social Security are not counted as government debt.

What would happen if the Social Security System invested in equities? The system would currently turn over less surplus to the Treasury, and the Treasury would have to issue more bonds to the public. Again, from the perspective of the government as a whole, this transaction amounts to a trade of bonds for equities with the public. Can such a trade benefit the public? Since equities are a claim on firms, government ownership of stock amounts to government ownership of some portion of the country's capital stock. So the preceding question can be rephrased. Does it matter who owns the capital stock? The analysis presented below attempts to shed light on that question. It turns out

that the policy of government investment in equities has either only minor or no effects on the government's budget and the saving rate of the economy. Whether the government's financing decisions have any economic effect depends on its ability to transfer risk across individuals. In the models considered in Section 3, that ability is absent, and hence government portfolio decisions are irrelevant. The overlapping generations model of Section 4 allows some scope for more efficient risk-sharing and the government's portfolio decision does affect economic behavior. Quantitatively, this effect turns out to be small.

3. A MODEL WITH INFINITELY LIVED AGENTS

In this section I use a model populated by infinitely lived agents (or more generally, the dynastic families possessing bequest motives as in Barro [1974]) to explore whether government investment in private capital affects the amount of tax revenue needed to support a given stream of transfer payments. Answering this question is analogous to answering the question of whether investment by the Social Security Administration in the stock market would have any impact on the financing of a given stream of Social Security payments. I analyze this question in a sequence of models that highlight the key issue, namely that equity premium considerations are unimportant and it is only the transferring of risk across generations that has any effect on economic outcomes.³ The model with infinitely lived agents clearly makes the point that when there is no possibility of transferring risk among agents, because all agents are essentially the same, the existence of an equity premium does not in any way allow government ownership of capital to influence economic outcomes.

To begin, I shall consider a world in which all transfers and taxes are lump sum. Private agents own some portion of the capital stock and the government owns the rest. The government may also issue debt. It finances transfer payments and the interest payments on debt through its earnings on capital and through taxes. I will show that in such a world the behavior of individuals is unaffected by the portion of the capital stock owned by the government. Essentially, any distribution of ownership of the capital stock is consistent with the initial path of transfers and taxes and has no effect on the consumption or saving decisions of individuals. In other words, the government's portfolio decision is irrelevant. I shall then extend this model to include distortionary taxes and show that the results are unchanged.

The Model with Lump Sum Taxes

This model economy is populated by people who live forever or, more generally, by the dynastic families in Barro (1974). Output is stochastic and is

³ For a detailed analysis of these issues, see Bohn (1997a, b).

produced via a standard neoclassical production function using capital and labor. The government finances lump sum transfers through lump sum taxes, the issuance of debt, and the return from its ownership of capital.

Individual Decisions

To start the analysis, consider the problem of the individual agent who wishes to maximize lifetime well-being or utility subject to a budget constraint. The individual owns some capital that earns $\rho(s_t)$ in state s at time t . That is, the return to capital is stochastic and, while one observes the actual return in any given period, future returns are uncertain and depend on the state of the economy in that period. Individuals also receive transfer payments from the government $Tr(s_t)$ and pay taxes $T(s_t)$. These transfers and taxes may, but need not, depend on the state of the economy. Individuals also own government bonds, $b(s_t)$, that pay $r(s_t)$ units of consumption in all states in period $t + 1$. Finally, given a capital stock at the beginning of period t , agents choose how much capital to bring into the next period, $k(s_t)$, and how much to consume this period, $c(s_t)$.

Formally, the representative agent maximizes discounted expected lifetime utility

$$\max_{t, S^t} \sum \beta^t u[c(s^t)] \pi(s^t)$$

subject to per-period budget constraints in each possible state s_t .

$$c(s^t) + b^d(s^t) + k(s^t) \leq w(s^t)n + \rho(s^t)k(s^{t-1}) + [1 + r(s^{t-1})]b(s^{t-1}) + Tr(s^t) - T(s^t),$$

where w is the real wage rate, n is exogenous labor supply, and ρ is the rate of return on capital. For simplicity, I assume that capital fully depreciates each period. Thus, agents are maximizing their utility, taking into account expectations of all possible future events. In the notation above, s_t is the realization of one of finitely many states of the economy at time t . s^t represents a particular history of realizations up to time t . That is, $s^t = (s_0, s_1, \dots, s_t)$ is a particular history of events up to time t . The set S^t represents all the possible histories that can occur. Each event occurs with probability $\pi(s_t)$ and each history occurs with probability $\pi(s^t)$. Each agent rents out labor and capital to firms in competitive rental markets and earns the appropriate marginal product of each factor.

The first-order conditions for optimal bond and capital accumulation are

$$u'[c(s^t)] = \beta \sum_{s^{t+1}} \pi(s^{t+1}) u'[c(s^{t+1})] [1 + r(s^t)] \quad (1a)$$

and

$$u'[c(s^t)] = \beta \sum_{s^{t+1}} \pi(s^{t+1}) u'[c(s^{t+1})] \rho(s^{t+1}). \quad (1b)$$

These conditions imply that agents accumulate assets so that they are just indifferent between consuming an extra unit of consumption in any particular state s_t or investing in either another bond or an extra unit of capital and consuming the proceeds of that investment next period. Also, since a government bond returns the same amount in each state at $t + 1$, it is less risky than holding capital whose return is uncertain. The interest on a government bond will, therefore, generally be less than the expected return on capital. That is, capital will on average earn a premium over bonds with the amount of the premium depending on the agent's aversion to risk and the underlying riskiness of the return on capital. It is this feature of bonds and capital that initially seems to suggest that the government, by issuing bonds and owning some more capital, can reduce the tax burden associated with any stream of transfer payments. However, as the first-order conditions make clear, both of these choices have the same value when adjusted for risk, namely the current marginal utility of consumption. Thus, there is no free lunch.

The Government

Each period the government makes some transfers, collects some taxes, and adjusts its portfolio by either issuing or repurchasing some government bonds or buying or selling some capital, x (or claims to the capital, which amount to the same thing). In each state, the government's net holding of assets obeys

$$b^s(s^t) - x(s^t) = b(s^{t-1})[1 + r(s^{t-1})] + Tr(s^t) - T(s^t) - \rho(s^t)x(s^{t-1}). \quad (2)$$

It is clear from this expression that, all other things equal, an increase in the capital stock held by the government at time $t - 1$ reduces the taxes that are necessary to maintain the same net asset position. The experiment we are interested in, however, is not what happens if someone donates an extra unit of capital to the government but what happens when the government increases its holdings of capital by issuing additional debt.

Market Clearing

For any allocation of consumption, bonds, and capital to be an equilibrium, it must be consistent with the resource constraints of the economy and with supply equaling demand. In particular, for each state the following equations hold:

$$c(s^t) + k(s^t) + x(s^t) = A(s_t)[k(s^{t-1}) + x(s^{t-1})]^\alpha n^{1-\alpha} \quad (3)$$

and

$$b^s(s^t) = b^d(s^t). \quad (4)$$

Equation (3) indicates that the amount consumed and invested must equal the output produced in the current period, and equation (4) requires that the supply

of bonds issued by the government must be equal to the demand for these bonds by the public.

Solution

The consumption decision of agents will now be shown to be independent of portfolio decisions of the government. Alternatively, agents do not care who owns the capital stock since they are indifferent between holding an extra unit of government debt or an extra unit of capital. In particular, consumption in any state is given by

$$c(s^t) = (1 - \beta)A(s_t)K(s^{t-1}), \quad (5)$$

where K is the aggregate capital stock equal to $k + x$. The accumulation of private capital is then expressed as

$$k(s^t) = \beta A(s_t)K(s^{t-1}) - x(s^{t-1}). \quad (6)$$

As long as government capital does not exceed $\beta A(s_t)K(s^{t-1})$, the above solutions satisfy the first-order conditions of agents and do not violate the economy's overall resource constraint. Thus, for any supportable path of taxes and transfer payments, individuals are indifferent as to who owns the capital stock.

The Model with Distortionary Taxes

Next consider the case where the government raises revenue through distortionary taxation. In this setting it is not so easy to represent analytically the solution to the decision problem of agents. However, by looking at the individual's first-order conditions and budget constraints along with the budget constraint and transversality condition of the government, one sees that the proportion of the capital stock owned by the government is irrelevant.

Individual Decisions

With distortionary taxes on both labor and interest income, the representative agent maximizes lifetime utility subject to the following per-period budget constraint,

$$c(s^t) + b^d(s^t) + k(s^t) \leq \rho(s^t)[1 - \tau(s^t)]k(s^{t-1}) + w(s^t)n[1 - \tau(s^t)] \quad (7) \\ \{1 + r(s^{t-1})[1 - \tau(s^t)]\}b(s^{t-1}) + Tr(s^t).$$

Unlike the previous budget constraint, the government now taxes wages and the return on capital and bonds at the rate τ . The first-order necessary conditions for optimal bond holdings and investment are

$$u'[c(s^t)] = \beta \sum_{s^{t+1}} \pi(s^{t+1})u'[c(s^{t+1})]\{1 + r(s^t)[1 - \tau(s^{t+1})]\} \quad (8a)$$

and

$$u'[c(s^t)] = \beta \sum_{s^{t+1}} \pi(s^{t+1}) u'[c(s^{t+1})] \{ \rho(s^{t+1}) [1 - \tau(s^{t+1})] \}. \quad (8b)$$

The consumer's accumulation of assets must also satisfy the transversality condition,

$$\lim_{j \rightarrow \infty} \sum_{j, s_{t+1}^{t+j} \in S_{t+1}^{t+j}} p(s_{t+1}^{t+j}) [k(s_t^{t+j}) + b(s_t^{t+j})] = 0, \quad (9)$$

where $p(s_t^{t+j}) = \beta^j \pi(s_t^{t+j}) \{ u'[c(s_t^{t+j})] / u'[c(s^t)] \}$ is the price of a contingent claim. In the above expression s_t^{t+j} indicates a particular history of states from t to $t+j$ and S_t^{t+j} is the set of all possible histories.

Government

The government's budget constraint is given by

$$b^s(s^t) - x(s^t) = b(s^{t-1}) \{ 1 + r(s^{t-1}) [1 - \tau(s^t)] \} + Tr(s^t) - \rho(s^t) \tau(s^t) k(s^{t-1}) - \tau(s^t) w(s^t) n - \rho(s^t) x(s^t) \quad (10)$$

and indicates that the government's net liability position depends on its debt, the net interest paid on that debt, its revenues from taxing income earned from capital and labor, as well as the revenue it earns on its own capital stock.⁴ The budget constraint implies that in states where capital has a relatively high rate of return, some debt is retired, while in states where capital's return is low, debt is issued. The government's net asset position must also satisfy the transversality condition

$$\lim_{j \rightarrow \infty} \sum_{j, s_{t+1}^{t+j} \in S_{t+1}^{t+j}} p(s_{t+1}^{t+j}) [b(s_t^{t+j}) - x(s_t^{t+j})] = 0. \quad (11)$$

Equilibrium

Formally, the definition of an equilibrium is given by

⁴ Using the above budget constraint and the first-order conditions of the representative agent, the government's lifetime budget constraint as of period t can be expressed as

$$\rho(s_t) \tau(s_t) k(s_{t-1}) + \rho(s_t) x(s_{t-1}) - b(s_{t-1}) = \sum_{j, s_{t+1}^{t+j} \in S_{t+1}^{t+j}} p(s_{t+1}^{t+j}) Tr(s_t^{t+j}) + \sum_{j, s_{t+1}^{t+j} \in S_{t+1}^{t+j}} p(s_{t+1}^{t+j}) \rho(s_t^{t+j}) [k(s_t^{t+j}) + x(s_t^{t+j})].$$

Notice that only the sum of private and government-owned capital stock enters the right-hand side of equation (11). Therefore, for any sequence of state-contingent prices, only the total capital stock and not its distribution affects the tax policies that are necessary to support a given stream of transfer payments.

Equilibrium: Given the initial conditions $b(s_{t-1}), x(s_{t-1}),$ and $k(s_{t-1}),$ an equilibrium is a sequence of quantities and prices $\{b(s), k(s), x(s), K(s), c(s), w(s), r(s), \rho(s), \tau(s), Tr(s)\}$ for all histories $s \in S_t^\infty$ satisfying the individual's first-order conditions (8a) and (8b), the individual's budget constraint (7), the government's budget constraint (10), the economy's resource constraint (3), and the transversality conditions of both the individual and the government (9) and (11).

Irrelevance Proposition:⁵ Suppose that $\{b(s), k(s), x(s), K(s), c(s), w(s), r(s), \rho(s), \tau(s), Tr(s)\}$ is an equilibrium, then any $\{\bar{b}(s), \bar{k}(s), \bar{x}(s), K(s), c(s), w(s), r(s), \rho(s), \tau(s), Tr(s)\}$ is an equilibrium if $\bar{b}(s), \bar{k}(s), \bar{x}(s)$ satisfy (a) $k(s), \bar{x}(s) \geq 0$ and $\bar{k}(s) + \bar{x}(s) = K(s),$ and (b) $\bar{b}(s)$ is defined recursively by (10).

Proof: The individual's first-order conditions and the economy's resource constraint are satisfied because the real allocations are identical in the two equilibria. The individual's transversality condition is, therefore, also satisfied. Equilibrium in the goods market and condition (b) imply that the household's budget constraint is also satisfied. Examining the lifetime budget constraint of the government from date t onward, one derives that

$$\begin{aligned} b(s^{t-1}) - x(s^{t-1}) &= \sum_{j=0}^T \sum_{s_i^{t+j} \in s_i^{t+j}} p(s_i^{t+j}) [Tr(s_i^{t+j}) - w(s_i^{t+j})\tau(s_i^{t+j})n \\ &\quad - \rho(s_i^{t+j})\tau(s_i^{t+j})K(s_i^{t+j})] \\ &\quad + \sum_{s_i^{t+T+1} \in s_i^{t+T+1}} p(s_i^{t+T+1}) [b(s_i^{t+T+1}) - x(s_i^{t+T+1})]. \end{aligned} \quad (12)$$

Because the first two terms are the same for both equilibria, the last term must be the same for both equilibria. Therefore, the transversality condition must hold for the second equilibrium. Hence, different distributions of the capital stock do not affect the aggregate capital stock, consumption, rates of return, tax rates, wages, or transfer payments.

As demonstrated in these models, the ownership of the capital stock has no effect on economic outcomes and is not an avenue that can be used to rescue the Social Security System in an economy where agents are altruistically linked to future generations and, hence, behave as if they were infinitely lived.

4. A MODEL WITH FINITE LIVED AGENTS

The previous two cases demonstrate that a premium in the return to capital relative to bonds is not sufficient for government portfolio decisions to have any

⁵ I would like to thank Andreas Hornstein for suggesting and helping me with this particular form of the argument.

real effect on consumer decisions. Changes in portfolio allocations do not affect the lifetime opportunities of the average individual, so they do not have any real consequence. In a model with finite lived agents, however, portfolio decisions generally will affect the economic behavior of consumers—not because capital earns a higher return than bonds but because a change in the portfolio of old agents must affect their consumption decisions. In the last period of life it is the only decision they have left to make. Thus, the government ownership of capital means that the current old agents hold more bonds. This consideration implies that their consumption stream has different risk characteristics than if the government owned no capital. Because the government's ownership of capital can transfer risk between current and future generations, it can change behavior.⁶ In the setting of infinitely lived agents, there is no one to whom they can transfer risk. But because we are now considering different generations, there is the potential for risk transfer. How big an effect policies involving portfolio composition may have is an open question. In this section, some rough estimates are formed in a simple two-period overlapping generations model. The results suggest that government ownership of capital may not be the boon that its proponents suggest.

The Individual

In the first period of life, a young individual works a fixed number of hours, n . With his earnings, he pays taxes, saves to finance consumption when old, and purchases goods for current consumption. Saving takes the form of ownership of the capital stock and government bonds. When the individual reaches old age, he receives transfers from the government, rental on the capital stock that then fully depreciates, and after-tax interest plus principal on his government bonds. With this income he purchases consumption goods. In this model economy production is stochastic and transfer payments are fixed. Formally, the individual's problem is

$$\max \{u[c^y(s^t)] + \beta \sum_{s^{t+1}} u[c^o(s^{t+1})]\pi(s^{t+1})\}$$

subject to the budget constraints

$$c^y(s^t) + k(s^t) + b(s^t) \leq w(s^t)n[1 - \tau(s^t)] + T^y \quad (13a)$$

and

$$c^o(s_{t+1}) \leq \rho(s_{t+1})[1 - \tau^k(s_{t+1})]k(s^t) + \{1 + r(s^t)[1 - \tau^k(s_{t+1})]\}b(s^t) + T^o, \quad (13b)$$

⁶ This idea is discussed in Volume II of the *Report of the 1994–1996 Advisory Council on Social Security* (Department of Health and Human Services 1997). The effects of government financing decisions on intergenerational risk-sharing are formally derived in Bohn (1997a, b), Smetters (1997), and Mariger (1997). Smetters shows that the risk-sharing engendered by the government's purchase of equities is equivalent to options contracts between generations.

where the last constraint must hold for each possible state s_{t+1} drawn from the set S_{t+1} . The superscripts y and o refer to young and old, respectively.

Here, as in the previous examples, s indexes the various possible states that can occur. The above specification assumes that agents know what state they are currently in but are unsure about next period's state. All they know is the probability, π , of any particular state occurring. Specifically, at time t , agents know how productive the economy is, the transfers that are given to both the current old and current young, the current tax rates on labor income, τ , and interest income, τ^k , the current wage rates, and the promised rate of interest on government bonds. They do not, however, know what these variables will be in the future. Thus, they attempt to maximize not only the utility from current consumption but expected utility from future consumption.

The first-order conditions for the problem are

$$u'[c^y(s^t)] = \lambda^y(s^t), \quad (14a)$$

$$\beta\pi(s_{t+1})u'[c^o(s_{t+1})] = \lambda^o(s_{t+1}) \text{ for each } s_{t+1} \in S_{t+1}, \quad (14b)$$

$$u'[c^y(s^t)] = \beta \sum_{s^{t+1}} u'[c^o(s^{t+1})]\pi(s^{t+1})\rho(s^{t+1})[1 - \tau^k(s^{t+1})], \quad (14c)$$

and

$$u'[c^y(s^t)] = \beta \sum_{s^{t+1}} \{1 + r(s^t)[1 - \tau^k(s^{t+1})]\} u'[c^o(s^{t+1})]\pi(s^{t+1}), \quad (14d)$$

where a prime indicates the first derivative and $\lambda^y(s^t)$ and $\lambda^o(s^{t+1})$ are the multipliers associated with the constraints (13a) and (13b). The last two constraints give the efficient consumption-saving decisions of the current young. These conditions state that at an optimum the marginal utility of forgoing one unit of consumption today must be equal to expected marginal utility of additional consumption tomorrow earned from the proceeds of investing in another unit of either capital or bonds. Notice that the last two equations also imply that the certain yield on a bond and the expected after-tax yield on capital must be such that the agent is indifferent between holding a bond or capital. As before, because the return on capital is uncertain, the premium that capital earns over bonds depends on the agent's degree of risk aversion.

Firms

Firms produce output by employing the labor of the young and renting capital from the old and from the government. The production function is constant returns to scale and is given by

$$Y(s^t) = A(s_t)K(s^{t-1})^\alpha n^{1-\alpha}, \quad (15)$$

where Y is aggregate per capita output, and K is the aggregate per capita capital stock. The maximization of profits implies that each factor receives its marginal product, which will depend on the productivity shock $A(s_t)$.

Government

The government issues bonds and purchases capital. It also supplies transfers to the young, T^y , and the old, T^o . These latter transfers may be thought of as Social Security although in reality the old receive more than just OASI payments alone. The government also raises revenue by taxing wage and capital income as well as the interest earned on bonds. Specifically, the government's budget constraint is

$$B(s^t) - x(s^t) = \{1 + r(s^{t-1})[1 - \tau^k(s^t)]\}B(s^{t-1}) - \rho(s^t)x(s^{t-1}) \quad (16) \\ + T^y + T^o - \tau(s^t)w(s^t)n - \tau^k(s^t)\rho(s^t)k(s^{t-1}),$$

where $B(s)$ is the per capita aggregate supply of government bonds and $x(s)$ is the per capita capital stock owned by the government. The government's net indebtedness $B - x$ is positively influenced by its repayment of existing debt, the interest on that debt, and transfer payments. The government's earnings on its capital stock, as well as the revenue from the taxation of labor, bonds, and the private sector's return on capital, all reduce the government's indebtedness.

Equilibrium

Equilibrium in this model is defined as a sequence of quantities (consumption, capital, and bond allocations), factor prices (wages, interest rates, and rental rates), and taxes and transfers that are consistent with each agent's maximization of expected utility, and the firms' maximization of profits. Equilibrium satisfies the individual's budget constraints (13a) and (13b), the government's budget constraint (16), and the government's transversality condition and results in the clearing of both the bond and goods markets. In particular for each possible history,

$$Y(s^t) = c^o(s^t) + c^y(s^t) + K(s^t) \quad (17a)$$

and

$$B(s^t) = b(s^t). \quad (17b)$$

Also, the per capita capital stock must equal its individual components, i.e., $K(s) = k(s) + x(s)$.

Unlike the case where agents are in effect infinitely lived, a similar irrelevance proposition does not apply. In the overlapping generations model, two separate budget constraints, one for the current old, (13b), and one for the current young, (13a), must hold simultaneously. Notice that the sum of these two budget constraints is the same as the budget constraint for the

infinitely lived agent. Thus any allocation that satisfies the economy's resource constraints and the government's budget constraint will satisfy the sum of the two agents' budget constraints; hence, total consumption will be unchanged. However, this allocation will not generally satisfy each budget constraint separately, and individual consumption will vary with changes in the distribution of capital. The variation in individual consumption implies that rates of return will have to change as well and that the same sequence of tax rates cannot support an identical path of transfer payments.

Analyzing the Effects of Government Ownership of Capital

To analyze the effects of government ownership of capital, I analyze the effect on average tax rates of changes in the proportion of the capital stock owned by the government. In doing so, the pattern of transfer payments and the government's net asset position, $B - x$, are fixed. As a result the experiment does not create any additional government indebtedness and maintains the level of benefits received by the elderly. The results of this experiment are suggestive but not definitive. The model I use is admittedly stylized. Moreover, I do not investigate plausible alternative fiscal policies, including those fixing the net present discounted value of government liabilities rather than fixing them in each and every period. The latter policy would produce a smoother stream of taxes than the one analyzed here but would be computationally much harder to implement. Also, because of the assumption that people live for two periods only, the benefits of risk-sharing are likely to be overemphasized in this framework. Old agents are required to hold all of the capital stock; thus any ownership of capital by the government reduces their exposure to rate-of-return risk. If the model included more periods, old agents could shift some of this burden to agents in their middle ages and thus reduce the risk-sharing benefits that ensue from the government's ownership of capital. The model also excludes other forms of risk-sharing arrangements, such as capital-gains loss-offsets and progressive taxation. Adding these features to the model would further reduce the gains to intergenerational risk-sharing.⁷

The equations used to solve the model include one that specifies the policy of fixing the government's net indebtedness and an equation that specifies the taxation of labor income relative to interest income. Equations 13(a,b), 14(c,d), 15, 16, and the two first-order conditions that determine the marginal product of capital and labor are also employed. Together with a behavioral relationship that specifies the government's purchase of capital, the solution to the model involves solving 11 independent equations in 11 unknowns. The variables solved for are the privately held capital stock, the publicly held capital stock, govern-

⁷I wish to thank Douglas Diamond and Kent Smetters for bringing these points to my attention.

ment bond issue, consumption by the young, consumption by the old, output, the interest rate paid on bonds, the rental rate on capital, wages, and the tax rates on labor and interest income, respectively. This system can be reduced to three equations that determine the interest rate, the aggregate capital stock, and the tax rate. In deriving these equations, I assume that the government maintains ownership of a fixed percentage of the capital stock, μ . It is also assumed that utility displays constant relative risk aversion and takes the form $u(c) = \frac{c^{1-\sigma}-1}{1-\sigma}$. Thus, the solution to this three-equation system yields the policy function for $K(s^t) = h_k[K(s^{t-1}), A(s_t), A(s_{t-1})]$, the functions $\tau(s^t) = h_\tau[K(s^{t-1}), A(s_t), A(s_{t-1})]$, and $r(s^t) = h_r[K(s^{t-1}), A(s_t), A(s_{t-1})]$.

To analyze the effects of government investment in capital, two slightly different models are simulated, one in which only labor is taxed, $\tau^k = 0$, and one in which all income is taxed at the same rate, $\tau = \tau^k$. For given values of transfers and net government indebtedness, I then compare tax rates and the aggregate capital stock in model economies in which the government owns 0, 2.5, 5, and 10 percent of the capital stock. The proposal of investing up to 40 percent of the Social Security Trust Fund in equities would result in a much smaller proportion of government ownership of the capital stock than any of the percentages considered. In 1995 the value of the Social Security Trust Fund was approximately \$458 billion, while the value of traded equity was greater than \$7.7 trillion. Thus, the experiments will, on this dimension, overstate the effects of the current proposal. In essence, I am comparing the equilibrium outcomes of four different economies. Transitional questions are, therefore, not addressed by this experiment.

Calibration

In calibrating the model, I envision a period as corresponding to 25 years. β is set at 0.5, which corresponds to an annual discount factor of roughly 0.973. Labor's share of output, α , is $2/3$, and the coefficient of relative risk aversion, σ , is set at 10, implying an average equity premium between 5.7 percent and 7.1 percent. Transfers to the old generation are set to equal 4 percent of steady-state output in the model. When only labor is taxed, such transfers are equal to the actual percentage of output distributed by OASI. The government's indebtedness is 1 percent of output and transfers to the young are roughly 2.5 percent of output, implying a steady-state tax rate on labor of 10.67 percent. This tax rate is close to the current tax rate of 10.52 percent on the OASI portion of the Social Security tax. Thus, the labor-tax-only model is calibrated to approximate the tax rate and the transfers that actually occur. Allowing the government also to tax capital increases the tax base and results in a lower steady-state tax rate and a somewhat higher level of capital and more output. The fraction of output transferred to the old is, therefore, also somewhat lower at 3.65 percent, although the old are receiving the same transfer in both models.

To analyze the effect on the average tax rate of government ownership of capital, I simulate both model economies over four generations or periods 1,000 times and take averages of the tax rates and capital stock that are produced by the simulations. Each simulation is started at capital's nonstochastic steady state, which is invariant to the government's portfolio allocation, and each succeeding capital stock is solved for based on the preceding realized value of capital and the past technology shocks. The tax rates and interest rate that are consistent with this solution are also obtained. The stochastic process for technology is identically and independently distributed with mean 1 and standard deviation of 0.08. The standard deviation was chosen to match the standard deviation of 25-year cumulative deviations from trend over the post-World War II period. This figure would represent the standard deviation of any generation's income from trend income. The standard deviation of this cumulative deviation from trend output was 0.13. I then used a standard deviation that was as close to 0.13 as possible and that still allowed for well-behaved policy functions of the capital stock.⁸ Because of the positive comovement of inputs with the technology shock, 0.13 is an upper bound on the variation in the technology shock. For example, Christiano and Eichenbaum (1992) obtain estimates of the relative variability of the technology shock to output anywhere from 48 to 90 percent. Therefore, 0.08 may not be an unreasonable number.

Results

The results of this experiment are reported in Tables 1 and 2. Table 1 includes the results when only labor is taxed, and Table 2 contains the results when both labor and interest income are taxed. For the case when only labor is taxed, one sees that average tax rates fall from 0.1059 to 0.1041 as the government increases its ownership of capital from zero to 10 percent of the aggregate capital stock. At 2.5 percent ownership, the decline in the average tax rate needed to support the level of transfer payments is negligible. It follows that ownership of equities by the Social Security Trust Fund would have little effect on the viability of the Social Security System. Because the decline in tax rates is so small, the capital stock is only marginally higher under the policy of government ownership of capital. In short, this proposed policy has little economic effect. The case where all income is taxed at the same rate is qualitatively similar. Basically, each economy's performance is not influenced by government portfolio decisions.

⁸ The models investigated above possess two steady states. One steady state, which is unstable, occurs at relatively low values of the capital stock. If the technology shock is too large, the capital stock potentially can enter this unstable region and the policy functions diverge.

**Table 1 Effects of Government Ownership of Capital
(only labor is taxed)**

Fraction of capital owned	0	2.5	5	10
Average tax rate	0.1059	0.1054	0.1049	0.1041
Standard deviation of tax rate	0.0074	0.0082	0.0089	0.0105
Average capital stock	0.1059	0.1061	0.1063	0.1066
Standard deviation of capital stock	0.0139	0.0141	0.0143	0.0147

**Table 2 Effects of Government Ownership of Capital
(all income is taxed)**

Fraction of capital owned	0	2.5	5	10
Average tax rate	0.0610	0.0606	0.0603	0.0596
Standard deviation of the tax rate	0.0042	0.0048	0.0053	0.0064
Average capital stock	0.1420	0.1421	0.1422	0.1425
Standard deviation of the capital stock	0.0163	0.0165	0.0166	0.0169

5. CONCLUSIONS

Current proposals for modifying Social Security have one key feature in common: namely, investing part of the trust fund in equities. Advocates believe that such a reallocation of the trust fund's portfolio will make the system more viable, and maintain the level of benefits without resorting to large increases in taxes. After analyzing the effects of such reallocation in some basic economic models, the results are not encouraging. Even though capital on average earns a higher rate of return than bonds, the government is not able to take much advantage of this differential, because only the ability to shift risk matters. The results in this regard are similar to those found in Bohn (1997a, b), Mariger (1997), and Smetters (1997). Quantitatively, this risk shifting from old to young does not significantly affect the government's budget or the economic behavior of individuals. In short, under the fiscal policies studied above, there is not much to be gained by government ownership of the capital stock. Actuarial soundness of the Social Security System will have to be achieved through other means.

REFERENCES

- Barro, Robert J. "Are Government Bonds Net Wealth?" *Journal of Political Economy*, vol. 82 (November/December 1974), pp. 1095–1117.
- Bohn, Henning. "Risk Sharing in a Stochastic Overlapping Generations Economy." Manuscript. June 1997a.
- . "Social Security Reform and Financial Markets." Manuscript. June 1997b.
- Christiano, Lawrence J., and Martin Eichenbaum. "Current Real-Business-Cycle Theories and Aggregate Labor-Market Fluctuations," *American Economic Review*, vol. 82 (June 1992), pp. 430–50.
- Department of Health and Human Services. *Report of the 1994–1996 Advisory Council on Social Security*. Washington: Government Printing Office, January 1997.
- Donlan, Thomas G. "Social Investment: Reform of Social Security Requires Private Investment and More," *Barron's Online* (<http://www.barrons.com>), January 31, 1997.
- Greenspan, Alan. Remarks at the Abraham Lincoln Award Ceremony of the Union League of Philadelphia. December 6, 1996.
- Marcks, Ronald H. "Social Security's Most Basic Infirmity," *Wall Street Journal*, January 16, 1997.
- Mariger, Randall P. "Social Security Privatization: What It Can and Cannot Accomplish," Finance and Economics Discussion Series, No. 1997–32. Washington: Board of Governors of the Federal Reserve System, Divisions of Research & Statistics and Monetary Affairs, June 1997.
- Miron, Jeffrey A., and David N. Weil. "The Genesis and Evolution of Social Security," *NBER Working Paper* 5949. March 1997.
- Munnell, Alicia H. *The Future of Social Security*. Washington: Brookings Institution, 1977.
- Smetters, Kent. "Investing the Social Security Trust Fund into Equity: An Options Pricing Approach." Washington: Congressional Budget Office, July 1997.
- Stein, Herbert. "Social Security and the Single Investor," *Wall Street Journal*, February 5, 1997.
- U.S. General Accounting Office. Report to the U.S. Senate Committee on Finance and the House of Representatives Committee on Ways and Means, *Social Security Administration: Significant Challenges Await New Commissioners*. Washington: Government Printing Office, February 1997.

Fisher and Wicksell on the Quantity Theory

Thomas M. Humphrey

The quantity theory of money, dating back at least to the mid-sixteenth-century Spanish Scholastic writers of the Salamanca School, is one of the oldest theories in economics. Modern students know it as the proposition stating that an exogenously given one-time change in the stock of money has no lasting effect on real variables but leads ultimately to a proportionate change in the money price of goods. More simply, it declares that, all else being equal, money's value or purchasing power varies inversely with its quantity.

There is nothing mysterious about the quantity theory. Classical and neo-classical economists never tired of stressing that it is but an application of the ordinary theory of demand and supply to money. Demand-and-supply theory, of course, predicts that a good's equilibrium value, or market price, will fall as the good becomes more abundant relative to the demand for it. In the same way, the quantity theory predicts that an increase in the nominal supply of money will, given the real demand for it, lower the value of each unit of money in terms of the goods it commands. Since the inverse of the general price level measures money's value in terms of goods, general prices must rise.

In the late nineteenth and early twentieth centuries, two versions of the theory competed. One, advanced by the American economist Irving Fisher (1867–1947), treated the theory as a complete and self-contained explanation of the price level. The other, propounded by the Swedish economist Knut Wicksell (1851–1926), saw it as part of a broader model in which the difference, or spread, between market and natural rates of interest jointly determine bank money and price level changes.

■ For helpful comments, thanks go to Mike Dotsey, Alice Felmler, Bob Hetzel, Rowena Johnson, Elaine Mandaleris, Ben McCallum, Ned Prescott, and Alex Wolman. The views expressed are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

The contrasts between the two approaches could hardly have been more pronounced. Fisher's version was consistently quantity theoretic throughout and indeed focused explicitly on the received classical propositions of neutrality, equiproportionality, money-to-price causality, and independence of money supply and demand. By contrast, Wicksell's version contained certain elements seemingly at odds with the theory. These included (1) a real shock explanation of monetary and price movements, (2) the complete absence of money (currency) in the hypothetical extreme case of a pure credit economy, and (3) the identity between deposit supply and demand at all price levels in that same pure credit case rendering prices indeterminate.

Despite these anomalies, Wicksell was able to derive from his analysis essentially the same conclusion Fisher reached. Both concluded that the monetary authority bears the ultimate responsibility for price level stability, a responsibility it fulfills either by determining some nominal variable—such as dollar price of gold, monetary base, bank reserves—under its control or by adjusting its lending rate in response to price level deviations from target.

The story of how Fisher and Wicksell reached identical policy conclusions from seemingly distinct models is instructive. It reveals that models appearing to be substantially different may be only superficially so. In the case of Fisher and Wicksell, it reveals that their models may not have been as dissimilar as often thought. Indeed, the alleged non-quantity-theory elements in Wicksell's work prove, upon careful examination, to be entirely consistent with the theory. In an effort to document these assertions and to establish Wicksell's position in the front rank of neoclassical quantity theorists with Fisher, the paragraphs below identify the two men's contributions to the theory and show how their policy conclusions derived from it.

1. FISHER'S VERSION OF THE QUANTITY THEORY

In his 1911 book *The Purchasing Power of Money*, Fisher gave the quantity theory, as inherited from his classical and pre-classical predecessors, its definitive modern formulation. In so doing, he accomplished two tasks. First, he expressed the theory rigorously in a form amenable to empirical measurement and verification. Indeed, he himself fitted the theory with statistical data series, many of them of his own construction, to demonstrate its predictive accuracy.

Second, he spelled out explicitly what was often merely implicit in the work of John Locke, David Hume, Richard Cantillon, David Ricardo, John Wheatley, and other early quantity theorists, namely the five interrelated propositions absolutely central to the theory. These referred to (1) equiproportionality of money and prices, (2) money-to-price causality, (3) short-run nonneutrality and long-run neutrality of money, (4) independence of money supply and demand, and (5) relative-price/absolute-price dichotomy attributing relative price

movements to real causes and absolute price movements to monetary causes in a stationary fully employed economy.¹

Fisher enunciated these propositions with the aid of the equation of exchange $P = (MV + M'V')/T$, which he attributed to Simon Newcomb even though Joseph Lang, Karl Rau, John Lubbock, and E. Levasseur had formulated it even earlier. Here P is the price level, M is the stock of hard or metallic money consisting of gold coin and convertible bank notes, V is the turnover velocity of circulation of that stock, M' is the stock of bank money consisting of demand deposits transferable by check, V' is its turnover velocity, and T is the physical volume of trade. Fisher's assumption that metallic money divides in fixed proportions between currency and bank reserves and that reserves are a fixed fraction of deposits allowed him to treat checkbook money as a constant multiple c of hard money. His assumption allows one to simplify his expression to $P = MV^*/T$, where $V^* = V + cV'$.

Of the equation's components, Fisher ([1911] 1963, p. 155) assumed that, in long-run equilibrium, the volume of trade is determined at its full-capacity level by real forces including the quantity and quality of the labor force, the size of the capital stock, and the level of technology. Save for transition adjustment periods in which the variables interact, these real forces and so the level of trade itself are independent of the other variables in the equation. Likewise, institutions and habits determine aggregate velocity, whose magnitude is fixed by the underlying velocity turnover rates of individual cashholders, each of whom has adjusted his turnover to suit his convenience (Fisher [1911] 1963, p. 152). Like the volume of trade, velocity is independent of the other variables in the equation of exchange. And with trade and velocity independent of each other and of everything else in the equation, it follows that equilibrium changes in the price level must be due to changes in the money stock.

Classical Propositions

All the fundamental classical quantity theory propositions follow from Fisher's demonstration. Regarding proportionality, he writes that "a change in the quantity of money must normally cause a proportional change in the price level" ([1911] 1963, p. 157). For, with trade and velocity independent of the money stock and fixed at their long-run equilibrium levels, it follows that a doubling of the money stock will double the price level.

Fisher realized, of course, that proportionality holds only for the *ceteris paribus* thought experiment in which trade and velocity are provisionally held fixed. In actual historical time, however, trade and velocity undergo secular changes of their own independent of the money stock. In that case, proportionality refers to the *partial* effect of money on prices. To this partial effect must

¹ For a discussion of these classical propositions, see Blaug (1995) and Patinkin (1995).

be added the parallel effects of coincidental changes in velocity and trade (see Niehans [1990], p. 277). The sum of these separate effects shows the influence of all on the price level. Thus if M , V^* , and T evolve secularly at the percentage rates of change denoted by the lowercase letters m , v^* , and t , respectively, then the price level P evolves at the percentage rate $p = m + v^* - t$. Fisher ([1911] 1963, pp. 246–47) himself expressed the matter precisely when he declared that the history of the price level is a history of the race between increases in the money stock and increases in the volume of trade.

Fisher was equally adamant on the neutrality of money other than during transition adjustment periods. Regarding long-run neutrality, he says that “An inflation of the currency cannot increase the product of . . . business” since the latter “depends on natural resources and technical conditions, not on the quantity of money” ([1911] 1963, p. 155). In short, trade’s long-run independence of money in the equation of exchange means that money cannot permanently influence real activity.

Money can, however, influence real activity temporarily. Indeed, the classical proposition regarding the short-run nonneutrality of money posits that very point. Fisher ([1911] 1963, pp. 58–72), in his theory of the cycle, attributes such nonneutrality to delays in the revision of lenders’ inflation expectations and the resulting sluggish adjustment of nominal interest rates. A monetary shock sets prices rising. Rising prices generate inflation expectations among business borrowers whose perceptions of current and likely future price changes are superior to those of lenders. These inflationary expectations engender corresponding expectations of higher business profits. Sluggish nominal loan rates, however, fail to rise enough to offset these rising expectations. Consequently, real loan rates fall. Spurred by the fall in real rates, business borrowers increase their real expenditure on factor inputs. Employment and output rise. Eventually, nominal loan rates catch up with and surpass business profit (and inflation) expectations. Real rates rise thereby precipitating a downturn.²

As for the proposition of unidirectional money-to-price causality, Fisher established it two ways. First, he denied that causation, under the gold standard then prevailing, could possibly run in the reverse direction from prices to money ([1911] 1963, pp. 169–71). To demonstrate as much, he supposed prices miraculously to double, the other variables in the exchange equation initially remaining unchanged. Far from inducing an accommodating expansion in the money stock, the price increase would, in an open trading economy, actually prompt that stock to contract. The stock would contract as the price increase, by rendering domestic goods expensive in relation to foreign ones, engendered

² As we will see, such nonneutralities are absent from Wicksell’s work. Adhering as he did to a real theory of the cycle, he denied that business fluctuations stem from monetary shocks (see Leijonhufvud [1997]). Such shocks in his view leave the economy always at full employment. Consequently, he held that neutrality of money prevailed in the short run as well as the long.

a trade balance deficit and a resulting external drain of monetary gold. The upshot is that the price increase would not cause a supporting rise in the money stock as reverse causation implies. Nor for that matter could the price increase spawn validating changes in the other variables of the exchange equation. The independence of those variables with respect to the price level rules out this possibility. In short, the price level is “the one absolutely passive element in the equation” (Fisher [1911] 1963, p. 172). Its movements are the result, not the cause, of prior changes in the quantity of money per unit of trade.

Alternatively, Fisher demonstrates M -to- P causality by showing that no variables in the exchange equation can intervene to absorb permanently the impact of a change in M and thus prevent the force of that impact from being transmitted to P . No compensating changes in trade will occur to blunt M 's impact since the two variables are independent in long-run equilibrium. Nor will M exhaust its effect in reducing velocity permanently. For cashholders have already established velocity at its desired level, a level independent of M .

Instead, Fisher ([1911] 1963, pp. 153–54) argued that money will transmit its full effect to prices through the following cash-balance adjustment mechanism. Let the money stock double from M to $2M$, the price level initially remaining unchanged. With prices and trade given, actual velocity $V^* = PT/2M$ falls to one-half the level cashholders desire it to be, or PT/M . In an effort to restore actual velocity to its desired level, cashholders will increase their rate of spending. The increased spending will, because trade is fixed at its full-capacity level, put upward pressure on prices. Prices will continue to rise until actual and desired velocities are the same ($V^* = 2PT/2M = PT/M$). At this point, prices will have doubled equiproportionally with money.

The remaining classical propositions follow directly from Fisher's analysis. Regarding the relative-price/absolute-price dichotomy, he denied that real factors change the absolute price level in a stationary, fully-employed economy. In particular, he insisted that price level changes cannot be caused by cost-push forces emanating from trade-union militancy, business-firm monopoly power, commodity shortages, and the like ([1911] 1963, pp. 179–80).³ Such forces, he says, drive relative prices, not absolute ones. In other words, given the money stock, velocity, and trade, real shock-induced changes in some relative prices produce compensating changes in others, leaving the absolute price level unchanged. Real shocks, if they are to affect absolute prices as well as relative ones, must somehow also cause changes in M , V^* , or T . Fisher saw little reason to expect them to do so. And even if they did, their effect would always be so

³ In his 1920 book *Stabilizing the Dollar*, Fisher listed 41 frequently cited nonmonetary causes of inflation and noted that “while some of them are important factors in raising particular prices, none of them . . . has been important in raising the general scale of prices” (p. 11). In his view “no explanation of a general rise in prices is sufficient which merely explains one price in terms of another price” (p. 14).

small as to be swamped by exogenous changes in money.

Finally, with respect to independence of money supply and demand, Fisher attempts to establish it by arguing that the money stock owes its determination to “influences outside the equation of exchange,” that is, to influences other than the trade-to-velocity ratio T/V^* ($= M/P$) which constitutes the public’s real demand for money ([1911] 1963, p. 90). For a closed gold-standard economy, these outside influences include the rate of gold production as influenced by new gold discoveries and technological innovations, both of which temporarily lower the metal’s production cost below its market value and so give a profit boost to mining. For open economies operating on the gold standard, additional external influences include foreign price levels. These, when high or low relative to the domestic price level, induce specie flows through the balance of payments. Such specie flows in turn raise or lower the domestic money stock and through it the domestic price level. From the viewpoint of the open domestic economy, money-stock changes are predetermined exogenously by the height of the foreign price level. These money-stock changes then endogenously affect domestic prices. As Fisher put it, “the price level outside of New York City . . . affects the price level in New York City only *via* changes in the money in New York City. Within New York City it is the money which influences the price level, and not the price level which influences the money” ([1911] 1963, p. 172).

Today, of course, we would say that an open economy’s money stock is endogenously determined by the requirement that domestic price levels move in step with foreign ones to maintain equilibrium in the balance of payments (see Friedman and Schwartz [1991], p. 42). But Fisher, by contrast, argued that the open economy’s money stock is determined exogenously by the *given* state of the balance of payments resulting from the *given* foreign (relative to domestic) price level.

We will see in Section 4 below, however, that he did correctly apply the exogeneity, or independence, proposition to so-called compensated dollar and inconvertible paper standard regimes. He recognized that, in such regimes, the policy authority governs money exogenously either through control of the gold weight of the dollar or through the high-powered monetary base consisting of the authority’s own liabilities. Through these instruments, the authority renders the money stock independent of money demand.

2. WICKSELL’S INTERPRETATION OF THE QUANTITY THEORY

Knut Wicksell’s perception of the classical quantity theory, as expounded in his 1898 *Interest and Prices* and Volume 2 of his 1906 *Lectures on Political Economy*, was less comprehensive than Fisher’s. Wicksell understood the

theory to mean only the proposition that prices are proportional to hard money, or metallic currency, in long-run equilibrium. This proportional relationship was, he believed, established through the operation of a real balance effect. In his view, cashholders had a well-developed demand for a constant stock of real cash balances. This demand together with the given nominal money supply ensured price level determinacy.

Thus a random shock to the price level that temporarily raised it above its equilibrium level would, by making actual real balances smaller than desired, induce cashholders either to cut their expenditure on or to increase their sales of goods in an effort to restore the desired level of real balances. The resulting excess supply of goods on the market would put downward pressure on prices until they reestablished their initial proportional relationship to the unchanged money stock, thus restoring real balances to equilibrium. In Wicksell's own words:

suppose that for some reason or other commodity prices rise while the stock of money remains unchanged The cash balances will gradually appear to be *too small in relation to the new level of prices* I therefore seek to enlarge my balance. This can only be done . . . through a *reduction* in my *demand* for goods and services, or through an *increase* in the *supply* of my own commodity . . . or through both together. The same is true of all other owners and consumers of commodities. But in fact nobody will succeed in realizing the object at which each is aiming—to increase his cash balance; for the sum of individual cash balances is limited by the amount of the available stock of money, or rather is identical with it. On the other hand, the universal reduction in demand and increase in supply of commodities will necessarily bring about a continuous fall in all prices. This can only cease when prices have fallen to the level at which the cash balances are regarded as *adequate*, [that is, when] prices . . . have fallen to their original level. ([1898] 1965, pp. 39–40)

This same stability condition, Wicksell noted, ensured that a decrease in the money stock would, by rendering real balances smaller than desired, induce a proportional fall in spending, and therefore prices, to restore real balances to their desired level. For Wicksell, then, the classical quantity theory implied money stock and price level proportionality achieved through real balance effects.

Pure Cash Economy

Wicksell found the theory to be perfectly valid for hypothetical pure cash economies in which no banks exist to issue checkable deposits, all transactions being mediated entirely by gold currency. In such economies, a demand for a fixed quantity of real gold balances ensures that prices move proportionally to money in long-run equilibrium. Thus newly discovered gold in a closed economy will, at initially unchanged prices, make real balances larger than

desired. Cashholders will spend the excess, thereby putting upward pressure on prices which rise proportionally to the increased monetary gold stock.

In an open trading economy, cashholders' adjustments will induce equilibrating real balance effects abroad as well as at home. For let all goods worldwide be tradeables—exportables and importables—whose prices are, by the law of one price, kept everywhere the same by the operation of commodity arbitrage. Then the increased home expenditure on these goods, induced by the gold discovery and resulting excess cash balance, will raise prices abroad thus eroding real balances there. In an effort to rebuild their balances, foreigners cut their spending on and increase their offer of tradeables. The resulting trade surplus is financed by a specie inflow that restores foreign real balances to their desired level. Real balance effects operate to establish proportionality between money and prices throughout the world (see Myhrman [1991], pp. 269–70).

Mixed Cash-Credit Economy

To Wicksell, however, the classical quantity theory, applying as it did to pure cash economies, seemed much too narrow and antiquated. It omitted banks and the deposit liabilities they issue by way of loan. It therefore could account neither for the influence of checking deposits on the price level, nor for how both variables move from one equilibrium level to another. Nor for that matter could it account for the forces inducing their movement. To overcome these deficiencies, Wicksell sought to supplement the quantity theory with a description of the mechanism through which monetary equilibrium is disturbed and subsequently restored in mixed cash-credit, or currency-deposit, economies. Thus was born his celebrated analysis of the cumulative process (see Jonung [1979], pp. 166–67, Laidler [1991], pp. 135–39, Leijonhufvud [1981], pp. 151–60, and Patinkin [1965], pp. 587–97).

That analysis attributes deposit and price level movements to discrepancies between two interest rates. One, the market or money rate, is the rate banks charge on loans and pay on deposits. The other is the natural or equilibrium rate that equates desired saving with intended investment at full employment and that also corresponds to the expected marginal yield or internal rate of return on newly created units of physical capital. Or, equivalently, it is the rate that equates aggregate demand for real output with the available supply.

When the loan rate lies below the natural rate such that the cost of capital is less than capital's expected rate of return, planned investment will exceed planned saving. Entrepreneur investors seeking to finance new capital projects will wish to borrow more from banks than savers deposit there. Since banks accommodate these extra loan demands by creating checking deposits, a deposit expansion occurs. This expansion, by underwriting the excess *desired* aggregate demand implicit in the investment-saving gap, transforms it into excess *effective* aggregate demand that spills over into the commodity market to put upward

pressure on prices. In so doing, the deposit expansion produces a persistent and cumulative rise in prices for as long as the interest differential lasts.

Now Wicksell argued that, in mixed cash-credit economies using currency and bank deposits convertible into currency, the rate differential would quickly vanish. The public's demand for real cash balances ensures as much. For let cashholders transact a certain portion of their real payments in currency. Then a rise in prices stemming from the rate differential necessitates additional currency to satisfy that real transaction demand. The ensuing public conversion of deposits into currency and the resulting drain on bank reserves induces banks to raise their loan rates until they (loan rates) equal the natural rate. This last step stems the reserve drain and also brings the price rise to a halt. If banks, because they initially possessed excess reserves, were willing to let reserves run down a bit, then prices would stabilize at the new, higher level. But if banks possessed no excess reserves and so had to restore reserves to their initial level following the price rise, then they (banks) would continue to raise the market rate above the natural rate until prices returned to their pre-existing level. Either way, a quantity theory element in the form of the public's demand for currency works to anchor the price level in the mixed cash-credit economy. Nominal determinacy prevails in that economy as it did in the pure cash economy.

Cumulative Process Model

Expressed symbolically and condensed into a simple algebraic model, Wicksell's cumulative process can be put through its paces to reveal the exact workings of its constituent quantity theory elements. Since these elements have provoked so much controversy in the Wicksell literature, it is important to specify precisely how Wicksell used them.⁴ Assume with Wicksell that all saving is deposited with banks, that all investment is bank-financed, that banks lend solely to finance investment, and that full employment prevails such that shifts in aggregate demand affect prices but not real output. Then his model reduces to the following equations linking the variables investment I , saving S (both planned, or *ex ante*, magnitudes), loan rate i , natural rate r , loan demand L_D , loan supply L_S , excess aggregate demand E , change in the stock of checkable deposits dD/dt , price level change dP/dt , and market-rate change di/dt .

The first equation says that planned investment exceeds saving when the loan rate of interest falls below its natural equilibrium level (the level that equilibrates saving and investment):

$$I - S = a(r - i), \quad (1)$$

⁴ For similar attempts to model algebraically the cumulative process see Brems (1986), Eagly (1974), Frisch (1952), Laidler (1975), Niehans (1990), and Uhr (1960).

where the coefficient a relates the investment-saving gap to the interest differential that creates it.

The second equation states that the excess of investment over saving equals the additional checkable deposits newly created to finance it,

$$dD/dt = I - S. \quad (2)$$

In other words, since banks create new checkable deposits by way of loan, deposit expansion occurs when banks lend to investors more than they (banks) receive from savers. Thus equation (2) admits of the following derivation. Denote the investment demand for loans as $L_D = I(i)$, where $I(i)$ is the schedule relating desired investment spending to the loan rate of interest. Similarly, denote loan supply as the sum of saving plus new deposits created by banks in accommodating loan demands. In short, $L_S = S(i) + dD/dt$. Equating loan demand and supply and solving for the resulting gap between investment and saving yields equation (2).

The third equation says that the new deposits, being spent immediately, spill over into the commodity market to underwrite the excess aggregate demand for goods E implied by the gap between investment and saving:

$$dD/dt = E. \quad (3)$$

The fourth equation says that this excess aggregate demand bids up prices, which rise in proportion to the excess demand:

$$dP/dt = bE, \quad (4)$$

where the coefficient b is the factor of proportionality between price level changes and excess demand.

Substituting equations (1), (2), and (3) into (4), and (1) into (2), one obtains

$$dP/dt = ab(r - i) \quad (5)$$

and

$$dD/dt = a(r - i), \quad (6)$$

which together state that price inflation and the deposit growth that underlies it stem from the discrepancy between the natural and market rates of interest.

Finally, since bankers must at some point raise their loan rates to protect their gold reserves from inflation-induced cash drains into hand-to-hand circulation, one last equation,

$$di/dt = g dP/dt, \quad (7)$$

closes the model. This equation says that bankers, having worked off excess reserves, now raise their rates in proportion to the rate of price change (g being the factor of proportionality). The equation ensures that the loan rate

eventually converges to its natural equilibrium level, as can be seen by substituting equation (5) into the above formula to obtain

$$di/dt = gab(r - i). \quad (8)$$

Solving this equation for the time path of the loan rate i yields

$$i(t) = (i_0 - r)e^{-gabt} + r, \quad (9)$$

where t is time, e is the base of the natural logarithm system, i_0 is the initial disequilibrium level of the loan rate, and r is the given natural rate. With the passage of time, the first term on the right-hand side vanishes and the loan rate converges to the natural rate. At this point, monetary equilibrium is restored. Saving equals investment, excess demand disappears, deposit expansion ceases, and prices stabilize at their new, higher level.⁵

3. WAS WICKSELL A QUANTITY THEORIST?

At first glance the preceding model, especially equation (5), appears to attribute price level changes directly to the interest rate differential rather than to monetary causes. This point is sometimes cited as evidence that Wicksell was not a quantity theorist (see Greidanus [1932], p. 83, and Adarkar [1935], p. 27, as cited in Marget [1938], pp. 183, 187). But it is patently obvious that the model is perfectly consistent with the quantity theory when monetary shocks generate the rate differential. Under these conditions the differential and the resulting price movements clearly have a monetary origin.

Indeed, Wicksell himself described how a monetary impulse would trigger the cumulative process consistent with the classical quantity theory. Assuming the monetary impulse took the form of a gold inflow from abroad, he noted that the new gold ordinarily would be deposited in banks. So deposited, the gold would augment bank reserves beyond the level banks desired to hold. The resulting pressure of excess reserves would, he argued, induce banks to lower their loan rate below the natural rate, thus precipitating the cumulative rise in the volume of bank money (deposits) and prices. Under these conditions, one could confidently attribute changes in both the stock of deposits and the price level to preceding changes in the monetary gold stock.

Having recognized potential monetary origins of the cumulative process as a theoretical possibility, however, Wicksell rejected this possibility on empirical grounds. His study of nineteenth-century British prices and interest rates had

⁵ Of course if there were no excess reserves to begin with, prices would have to stabilize at their pre-existing level. Bankers, having no excess reserves to lose, would adjust their loan rates either to forestall all reserve drains or to reverse (annul) drains that had already occurred. Either way, prices would stabilize at their initial level.

convinced him that the cumulative process typically originated not in monetary shocks to the loan rate but rather in real shocks to the natural rate. His consequent stress on real shocks in the form of wars, technological progress, innovations, and the like has spurred some scholars to ask: if real shocks predominate over monetary shocks in generating the rate differential, doesn't it follow that the resulting price level movements are real rather than monetary phenomena, contrary to the quantity theory?

In answering this question in the affirmative, these scholars imply that Wicksell may have done more to subvert the theory than to support it. Thus Lars Jonung states:

Wicksell's approach emphasizes nonmonetary developments, that is "real" factors, as the principal sources of price changes. Although the monetary sector has a central position in the transmission mechanism from "real" developments to changes in prices, there is a tendency to ignore monetary factors in a theory that assumes that movements in the real rate are the driving force behind deflations and inflations. It is thus easy to end up with a theory of the price level that relates the behavior of prices directly to variables that influence the real rate, such as changes in the flow of innovations and technological improvements. Here Wicksell's theory has much in common with the Schumpeterian "longwave explanation," which associates price level changes with the introduction of new production techniques, which implies that non-monetary factors are the causes behind long-run changes in prices. (Jonung [1979], p. 179; see also Cagan [1965], p. 253, and Laidler [1997], p. 5)

What such interpretations overlook, however, is that Wicksell himself always saw his cumulative process model as embodying the quantity theory and being entirely consistent with it. His model was to him nothing less than a full-scale extension of the theory to account for the influence of bank deposits on the price level. In particular, his equations (3) and (4) upon substitution reduce to $dP/dt = b(dD/dt)$. In so doing, they reveal that a price level change could never occur without the accompanying change in the supply of deposits to support it.

In short, real shocks and the resulting rate differential alone could never sustain price level changes. Instead, something else is required to translate shocks into commodity price inflation. Something, in other words, must finance the excess demand for goods that keeps prices rising. That something is deposit expansion. Without it, excess demand and price increases could never occur and the cumulative process would be abortive. The upshot is that Wicksell thought the key factor underlying and permitting price movements was deposit expansion, not real shocks and rate differentials.

Of the few commentators who underscore this point, none are more emphatic than Charles Rist and Arthur Marget. Rist ([1938] 1966, p. 300) likens Wicksell to Voltaire's sorcerer, whose incantations could kill a herd of cattle if accompanied by a lethal dose of arsenic. In Wicksell's case, the arsenic—the

true cause—was an elastic supply of deposits. The incantations took the form of rate differentials. Similarly, Marget (1938, p. 183) cites “abundant passages in Wicksell’s writings which show that he did think of the ‘plentiful creation of money’ (that is, bank-credit, or the M' of our equation) as being the crucial link in the [cumulative] process.” In short, changes in the stock of deposits were to Wicksell the one absolutely necessary and sufficient condition for price level movements.

Critique of Tooke’s Interest Cost-Push Theory

Nowhere did Wicksell express this view more forcefully than in his famous critique of Thomas Tooke (Wicksell [1898] 1965, pp. 99–100, and [1906] 1978, pp. 180–87). Tooke, author of the celebrated *History of Prices* and leader of the English Banking School, had disputed, indeed scorned, the quantity theoretic doctrines of the rival Currency School. In opposition to those doctrines, Tooke, in his 1844 volume *An Inquiry into the Currency Principle* (Tooke [1844] 1959), argued that price level changes stem from cost-push forces originating in the real economy rather than from disturbances originating in the monetary sector. In particular, he argued that interest rate increases, by raising the cost of doing business, would raise general prices as the increased costs were passed on to buyers. The resulting price inflation, Tooke implied, would occur even in the face of a constant money stock.

Wicksell, however, maintained that such price level increases could never occur unless underwritten by expansion of that stock. According to him, it is deposit growth stemming from a two-rate differential, and not interest cost-push per se, that constitutes the necessary condition for general prices to rise. Without the accommodating monetary growth, the interest cost-push forces would, he insisted, exhaust themselves in changing relative, not absolute, prices ([1906] 1978, p. 180). The prices of interest-intensive goods would rise relative to the prices of non-interest-intensive ones. But the general price level would remain unchanged. For if the money stock were constant and banks possessed no excess reserves, any rise in the natural rate would force bankers to engineer a matching rise in the loan rate to protect their reserves from cash drains into hand-to-hand circulation. The two rates would remain equal and prices would stay constant. Only if banks initially possessed excess reserves could a positive shock to the natural rate permanently raise the equilibrium price level. And even here the price increase is attributable to the monetary factor—the excess reserve—that permits it to occur. All of which is consistent with the quantity theory and confirms Wicksell’s adherence to it.

Pure Credit Economy

To summarize, Wicksell had shown that the quantity theory applies perfectly to the pure cash economy. He had then shown that, when augmented to account

for the influence of deposit-financed demand on prices, it applies to mixed cash-credit economies as well. In both cases, he had established that a real currency demand together with an independent nominal currency supply are sufficient to pin down the price level. Seeking to extend the theory to its logical limit, he next applied it to the hypothetical extreme case of a pure credit economy in which no currency exists and all transactions are settled by transfers of deposits on the books of banks. Here he showed that the theory fails to hold in the absence of central bank intervention.

According to him, it fails to hold in the first place because the pure credit economy employs no currency to which the theory can apply. With currency absent, no demand for and supply of it exists to determine the price level. Nor can deposit demand and supply be relied upon to determine the price level. For, in the pure credit economy, the two deposit variables are identical to each other at all price levels. Being identical, they cannot exhibit demand-supply independence as price determinacy requires. Wicksell explains:

in our ideal [pure credit] state every payment . . . is accomplished by means of cheques or *giro* facilities. It is then no longer possible to refer to the supply of money as an independent magnitude, differing from the demand for money. No matter what amount of money may be demanded from the banks, that is the amount which they are in a position to lend The banks have merely to enter a figure in the borrower's account to represent a credit granted or a deposit created. When a cheque is then drawn and subsequently presented to the banks, they credit the amount of the owner of the cheque with a deposit of the appropriate amount (or reduce his debit by that amount). The "supply of money" is thus furnished by the demand itself. . . . It follows that . . . the banks can raise the general level of prices to any desired height. ([1898] 1965, pp. 110–11)

With deposit supply identical to demand at all prices, there is no unique equilibrium price level or deposit quantity. Rather, there is an infinity of price-quantity equilibria. The price level, in other words, is indeterminate. Wicksell's cumulative process model applied to the pure credit economy cannot determine it.

Instead, his credit economy model specifies the rate of rise of the price level dP/dt (see Leijonhufvud [1997], p. 8). Starting from some historically given position, this rise can continue indefinitely as long as a natural-rate/market-rate disparity persists, that is, as long as banks are under no reserve pressure to raise their rates. Since no currency demand exists to drain reserves in the pure credit economy, banks need hold no reserves other than central bank credit. And even this form of reserve is unnecessary in a banking system—Wicksell's "ideal" system—composed of a single central bank with branches in every town and hamlet (see Uhr [1960], p. 222). As a central bank, the ideal bank need hold no credit reserves with itself. Moreover, as a monopoly institution, the ideal bank can lose no reserves through the clearing house to other banks (of which there

are none) and so need hold no reserves whatsoever. The result is a system totally devoid of reserve constraints to anchor nominal variables. In such a system, deposit supply possesses potentially unlimited elasticity. Consequently prices, in addition to being indeterminate, theoretically can rise (or fall) forever.

Wicksell insisted, however, that it was up to the central bank to impose nominal determinacy in this case. The central bank could do so through control of the market rate. By adjusting the rate when prices threaten to rise or fall, the bank could close and reverse the rate differential. In so doing, the bank could maintain prices and the supporting volume of deposits at fixed, determinate levels. Here the central bank's obligation to impose price determinacy replaces the missing reserve constraint to force equilibrating rate adjustment. Nominal determinacy is preserved, consistent with the quantity theory. In this way, Wicksell ensures that at least one element of the theory survives even in the pure credit case.

4. POLICY REFORM PROPOSALS

The preceding remarks contend that Wicksell was, commentators' views to the contrary notwithstanding, every bit as much a quantity theorist as Fisher. Evidence reveals that he, like Fisher, understood and indeed enriched the theory's postulates.

But there is a simpler way to prove he and Fisher saw things much the same as far as the quantity theory was concerned. That way is to compare the policy views of the two. One can employ a simple litmus test: a person essentially is a quantity theorist if he believes the monetary authority can stabilize the price level through control, direct or indirect, of the stock of money or nominal purchasing power. Both Fisher and Wicksell pass this test with flying colors.

Both advocated price level stability, albeit for different reasons. Fisher thought such stability would smooth, if not eliminate completely, the business cycle. In so doing, it would alleviate the overuse (stress, strain, exhaustion) of labor and capital resources endured in business booms and the loss of output and employment suffered during depressions. By contrast, Wicksell thought price stability would stop the arbitrary and unjust redistribution of income and wealth that unanticipated inflation and deflation produce. In this way, it would prevent the loss in aggregate social welfare that occurs, because of diminishing marginal utility of income, when unanticipated price movements transfer real income from losers to gainers.

Both also advocated that price stability be achieved through feedback policy rules. In this connection, both devoted their best efforts to devising effective rules. Each writer proposed rules directing the monetary authority to adjust its policy instrument in corrective response to price level deviations from target. Such instrument adjustment would in turn produce a corresponding adjustment

in the money stock. This latter adjustment would act to stabilize prices. The money stock was of key importance here. Only by operating through it could instrument adjustment stabilize prices.

In Fisher's famous compensated dollar plan, the policy instrument is the gold content of the dollar, or official dollar price of gold (see Patinkin [1993]). The monetary authority adjusts this price in response to price level deviations from target. Since the price level, or dollar price of goods, is by definition the dollar price of gold times the world gold price of goods, the authority must offset movements in the gold price of goods with compensating adjustments in the dollar price of gold so as to keep the general price level constant.

Fisher made it clear, however, that his compensated dollar plan would operate on the price level through the money stock. It would do so by changing both the physical amount and the nominal valuation of the nation's stock of monetary gold. Thus when world gold inflation was raising the dollar price of goods, the American policy authority would lower the official buying and selling price of gold. Industry and the arts, finding gold less expensive, would therefore demand more of it. Consequently, part of the nation's gold stock would be diverted from monetary to nonmonetary uses (see Lawrence [1928], p. 432). The resulting shrinkage in the stock of monetary gold would lower the price level. In addition, the reduced official price of gold, by producing a corresponding reduction in the nominal value of physical gold reserves, would lessen the nominal volume of paper money issuable against such backing (see Patinkin [1993], p. 16). This reduced nominal issue too would put downward pressure on prices. In sum, whether through physical reduction or nominal revaluation, the monetary gold stock would shrink and so too would the quantity of money and level of prices it could support.

Later on, in the mid-1930s, Fisher (1935, p. 97) proposed another policy rule. It had the central bank adjusting, via open market operations, the monetary base in response to price deviations from target. In this case, the price level was the goal variable, the monetary base was the instrument, and the money stock was the intermediate variable. To minimize slippage between the base instrument and the money stock, Fisher advocated a system of 100 percent required reserves behind deposit money.

Although Wicksell's preferred policy instrument differed from Fisher's, his activist feedback rule followed exactly the same pattern as Fisher's. The authority would adjust its policy instrument, namely its lending rate, in response to price deviations from target. In Wicksell's own words (1919, p. 183, cited in Jonung [1979], p. 168), "the Riksbank's tool to keep the price level . . . constant is to be found exclusively in its interest rate policy, such that the Riksbank has to increase its rates as soon as the price level shows a tendency to rise and lower them, as soon as it shows a tendency to fall." Such rate adjustments would in turn produce corresponding corrective movements in the money

stock. These latter movements then would stabilize prices.⁶ Together, these propositions constitute what Howard S. Ellis, in his classic *German Monetary Theory: 1905–1933*, called Wicksell’s “central theorem,” namely the theorem “that bank rate controls the price-level through its effect on the amount of available purchasing power” (Ellis 1934, p. 304).

Thus if prices were rising, the central bank would raise the bank rate. The rise in the bank rate would close the gap between it and the natural rate. The closing of the gap would eliminate the differential between the investment demand for and saving supply of loanable funds. The elimination of that differential would arrest growth in the stock of deposits and bring price rises to a halt. Further raising of the bank rate would cause deflationary monetary contraction, thereby reversing the preceding inflationary price movement and restoring prices to target. Here is a classic quantity theoretic prescription for achieving price stability through monetary means. It is proof positive that Wicksell, like Fisher, was a bona fide quantity theorist.

5. CONCLUSION

What then remains of the alleged difference between Fisher’s and Wicksell’s interpretation of the quantity theory? Not much, in this observer’s opinion. Any existing difference seems superficial rather than substantive, more semantic than real. And it virtually vanishes once their policy reform proposals are taken into account.

Commentators typically claim that interest rates are the key to Wicksell’s analysis, whereas for Fisher the money stock is pivotal. They likewise claim that real shocks initiate the inflationary process in Wicksell’s model, whereas monetary shocks do so in Fisher’s. True enough. But these distinctions largely lose force when one realizes that both men saw changes in the stock of monetary purchasing power consisting of bank deposits and currency as the one absolutely indispensable and potentially controllable factor responsible for price level changes. Moreover, both regarded this stock as the crucial intermediate variable connecting policy instruments to price targets. Finally, both concluded that the monetary authority bears the ultimate responsibility for monetary and price level stability, a responsibility it discharges by giving some nominal variable under its control a stable, determinate value. In so doing, both enunciated the principle of nominal determinacy, the sine qua non of the quantity theory. These similarities would seem to outweigh any differences.

One reads Fisher and Wicksell today not so much to note the contrasts in their analytical models as to appreciate the brilliant, prescient, and imaginative

⁶ Uhr (1991, p. 94) notes that Wicksell believed that the application of his rule would prevent the price level from varying more than three percentage points above or below its target or base-year level.

ways they applied the quantity theory. In arguing for price stability achievable through monetary means, both were adherents of monetary policy in the classical quantity theory tradition. Their two treatments are complementary rather than competitive.

REFERENCES

- Adarkar, Bhalchandra P. *The Theory of Monetary Policy*. London: P. S. King & Son, 1935.
- Blaug, Mark. "Why is the Quantity Theory of Money the Oldest Surviving Theory in Economics?" in Mark Blaug, ed., *The Quantity Theory of Money from Locke to Keynes and Friedman*. Aldershot, England: Edward Elgar, 1995.
- Brems, Hans. *Pioneering Economic Theory, 1680–1980*. Baltimore: Johns Hopkins University Press, 1986.
- Cagan, Philip. *Determinants and Effects of Changes in the Stock of Money, 1875–1960*. New York: Columbia University Press, 1965.
- Eagly, Robert. *The Structure of Classical Economic Theory*. New York: Oxford University Press, 1974.
- Ellis, Howard S. *German Monetary Theory: 1905–1933*. Cambridge, Mass.: Harvard University Press, 1934.
- Fisher, Irving. *100% Money*. New York: Adelphi, 1935.
- _____. *Stabilizing the Dollar*. New York: Macmillan, 1920.
- _____. *The Purchasing Power of Money: Its Determination and Relation to Credit, Interest, and Crises*. New York: Macmillan, 1911, reprinted, New York: Augustus M. Kelley, 1963.
- Friedman, Milton, and Anna J. Schwartz. "Alternative Approaches to Analyzing Economic Data," *American Economic Review*, vol. 81 (March 1991), pp. 39–49.
- Frisch, Ragnar. "Frisch on Wicksell," in H. W. Spiegel, ed., *The Development of Economic Thought: Great Economists in Perspective*. New York: J. Wiley, 1952.
- Greidanus, Tjardus. *The Value of Money*. London: P. S. King & Son, 1932.
- Jonung, Lars. "Knut Wicksell and Gustav Cassel on Secular Movements in Prices," *Journal of Money, Credit, and Banking*, vol. 11 (May 1979), pp. 165–81.
- Laidler, David. "The Wicksell Connection, the Quantity Theory and Keynes." Unpublished paper. 1997.

- _____. *The Golden Age of the Quantity Theory: The Development of Neoclassical Monetary Economics, 1870–1914*. Princeton: Princeton University Press, 1991.
- _____. “On Wicksell’s Theory of Price Level Dynamics,” in his *Essays on Money and Inflation*. Chicago: University of Chicago Press, 1975.
- Lawrence, Joseph Stagg. *Stabilization of Prices: A Critical Study of the Various Plans Proposed for Stabilization*. New York: Macmillan, 1928.
- Leijonhufvud, Axel. “The Wicksellian Heritage.” Unpublished Discussion Paper No. 5. Università Degli Studi di Trento-Dipartimento di Economia, 1997.
- _____. “The Wicksell Connection: Variations on a Theme,” in his *Information and Coordination: Essays in Macroeconomic Theory*. New York: Oxford University Press, 1981.
- Margat, Arthur W. *The Theory of Prices: A Re-examination of the Central Problems of Monetary Theory*, Vol. 1. New York: Prentice-Hall, 1938.
- Myhrman, Johan. “The Monetary Economics of the Stockholm School,” in Lars Jonung, ed., *The Stockholm School of Economics Revisited*. Cambridge: Cambridge University Press, 1991.
- Niehans, Jürg. *A History of Economic Theory: Classic Contributions, 1720–1980*. Baltimore: Johns Hopkins University Press, 1990.
- Patinkin, Don. “Concluding Comments on the Quantity Theory,” in Mark Blaug, ed., *The Quantity Theory of Money from Locke to Keynes and Friedman*. Aldershot, England: Edward Elgar, 1995.
- _____. “Irving Fisher and His Compensated Dollar Plan,” *Federal Reserve Bank of Richmond Economic Quarterly*, vol. 79 (Summer 1993), pp. 1–33.
- _____. *Money, Interest, and Prices*, 2d ed. New York: Harper & Row, 1965.
- Rist, Charles. *History of Monetary and Credit Theory from John Law to the Present Day*, 1938, translated by Jane Degras, 1940, reprinted, New York: Augustus M. Kelley, 1966.
- Tooke, Thomas. *An Inquiry into the Currency Principle*, 1844. Reprinted, London: The London School of Economics and Political Science, 1959.
- Uhr, Carl G. “Knut Wicksell, Neoclassicist and Iconoclast,” in Bo Sandelin, ed., *The History of Swedish Economic Thought*. London and New York: Routledge, 1991.
- _____. *Economic Doctrines of Knut Wicksell*. Berkeley: University of California Press, 1960.

- Wicksell, Knut. *Lectures on Political Economy*, Vol. 2, "Money," 1906, translated by E. Classen, edited by Lionel Robbins. London: Routledge and Kegan Paul, 1935, reprinted, New York: Augustus M. Kelley, 1978.
- _____. *Interest and Prices*, 1898, translated by R. F. Kahn. London: Macmillan, 1936, reprinted, New York: Augustus M. Kelley, 1965.
- _____. "Riksbanken och privatbankerna. Förslag till reform av det svenska penningoch kreditväsendet," [The Riksbank and the commercial banks. A proposal to a reform of the Swedish monetary and credit system]. *Ekonomisk Tidskrift* (1919), part 2, pp. 177–88.