

Alvin L. Marty is professor of economics at the Center for the Study of Business and Government at Baruch College, City University of New York (C.U.N.Y.). Rob Dittmar, an economist at the Federal Reserve Bank of St. Louis, provided mathematical expertise that was very helpful in clarifying the Lucas paper. Turin Bali, a Ph.D. candidate at C.U.N.Y., ran Box-Cox tests on the constant and semi-elastic demand schedules. Dan Steiner provided research assistance.

The Welfare Cost of Inflation: A Critique of Bailey and Lucas

Alvin L. Marty

THE CONCEPT OF THE OPTIMAL MONEY STOCK

The concept of the optimal stock of money introduced by Milton Friedman (1969) has stimulated much discussion and controversy. More recently, Lucas (1994) has provided new estimates of the welfare gain for the American economy from setting the optimal level of real cash balances. Before turning to a critique of Lucas, let us provide a rationale for this concept of optimality.

To society, real balances are produced at zero marginal costs. To an individual, however, there is a cost to holding cash that bears no interest rather than bonds, which yield a money rate of interest. An individual would willingly incur this cost only if real balances produced services on the margin equal to the foregone interest. A lower interest rate reduces this cost and increases the total services provided by real cash balances. One estimate of the consumption the consumer would forego—to get this potential maximum gain from holding cash—is the area under the money demand schedule between the quantity of cash held at the lowest possible rate of interest (perhaps zero) and the smaller quantity that would be held at a higher interest rate.¹

Clearly any measure of this gain depends on the particular demand

schedule used. Lucas utilizes a double log schedule rather than the semi-log schedule used by Friedman (1969) and Martin Bailey (1956). The Lucas' double log schedule yields greater welfare gains since the level of cash balances always increases as the interest rate approaches zero. In contrast, the semi-log schedule implies a finite level of cash is held at a zero nominal interest. In a later section we will provide a fuller discussion of these schedules.

LUCAS' EXPOSITION

What reduction in consumption would compensate for and measure the gain in utility from a larger stock of cash? An exact measure of this gain is derived by Lucas in the case of both the double log (the constant elasticity schedule) and semi-log demand schedules for real cash balances. He assumes a time-separable constant-relative-risk-averse utility function with consumption and the ratio of real balances to consumption as arguments. This function is compatible with steady growth of a non-durable output. A representative agent maximizes it subject to a budget constraint.

Such an infinite horizon model, (see Ramsey [1927] and Sidrauski [1967]), implies that we need not be concerned with changes in the real rate of interest as we vary the growth of nominal money balances. Note the contrast to Mundell (1963) and Tobin (1965) models in which the accumulation of non-interest bearing real balances competes with physical capital as vehicles for the savings of finite-lived individuals. In their models, a faster growth of nominal money raises the actual and anticipated rates of price change. The capital loss on real balances spurs the accumulation of physical capital. A new steady state is reached: Capital and per capita output are higher and the real rate of interest is lower. In the Ramsey-type model used by Lucas, the real rate is tied to the fixed utility discount factor of the representative agent. When the rate of

¹ Strictly speaking, this welfare measure utilizes a compensated demand schedule, which keeps the consumer at the same level of demand.

price change is fully anticipated, the money rate of interest adjusts to give borrowers and lenders the same fixed real rate. With a constant real rate, the demand schedule for real balances represents alternative steady states with the nominal interest rate adjusting one-for-one with the rate of price change.

Lucas approximates exact measures of the gain in utility by the consumer surplus under both the constant and semi-log schedules. For the range of interest rates in U.S. history, the approximations are very close to the exact compensating variation using these demand functions.²

THE NATURE OF THE DEMAND SCHEDULE: BAILEY VS. LUCAS

The constant elasticity schedule is $M/P = Ar^{-\eta} y^\beta$, where M is nominal money, P is the price level, y is real income and r the nominal interest rate (money rate). The income elasticity, β , is taken as unity and the estimated interest elasticity, η , is 0.5.³ Although the demand for real balances increases without limit as the money rate approaches zero, the integral converges.

If other taxes distort (such as a tax on labor income), it is uncertain whether welfare improves if the authorities drive the money rate of interest towards zero. Other taxes would have to be raised to run a budget surplus and retire money at a rate equal to the real interest rate minus the output growth rate. Any welfare gain from higher real balances would have to be weighed against the welfare cost of increased taxes. Even in this case, however, Lucas estimated optimal money rate remains very close to zero.

As Lucas notes, Bailey (1956) approximated the exact welfare loss by the consumer surplus integral under the semi-log demand schedule, $M/P = e^\gamma e^{-aE}$, where e^γ is an index of real balances held at zero anticipated inflation.⁴ Lucas prefers a constant elasticity demand function on the grounds that it fits the American data.

Before proceeding with a critique of Lucas, it is useful to discuss the properties of the constant elasticity function. Unlike

the semi-log, the constant elasticity function does not generate a Laffer curve. Along a Laffer curve, as the inflation tax rate (the money rate) rises, the revenue increases. It reaches a maximum, then it declines. Using the Phelps (1973) and Auernheimer (1974) definition of the revenue as $(M/P)(r)$, the revenue ($R = rAr^{-\eta} = Ar^{(1-\eta)}$) continuously increases with inflation.⁵

Although the double log function may fit the American data that includes only moderate rates of inflation, it is not evident that the schedule should be extended—as Lucas does—to regions of hyperinflation, or for that matter, a deflation of prices approaching the real rate. Bali (1998) has run tests using the Box-Cox transformation to determine whether the constant or semi-elastic function best fits the data. Over the range of data in the United States (therefore not at hyperinflation or at rates of interest approaching zero), the double log performs better. For non-U.S. hyper-inflations, the semi-log fits better, however. We should be very cautious about extrapolating to non-observable ranges of the data. Since data are not available at rates of interest close to zero, we proceed to discuss Lucas' (1994) extension to hyperinflation where international data are available.

EXTENSION TO HYPERINFLATION

Cagan (1956) found that the semi-log fits the data for seven European countries during periods of hyperinflation. The semi-log generates a Laffer curve: The steady state maximum revenue is at

$$\frac{dM}{dt} \frac{1}{M} = \frac{1}{a}$$

Cagan concluded that governments often inflated beyond this point, an overshooting paradox which he explained as follows. The revenue is the product of real balances (which depend on expected inflation) and the tax rate (the actual inflation). The authorities, therefore, could exploit a lag in expected inflation to temporarily get more revenue when, as Cagan assumed,

² For the details of these calculations, see Lucas (1994).

³ It is crucial that the income elasticity is unity. If this were not the case, we could not use steady state analysis. If the income elasticity were less than one, real balances would fall as a ratio to real income and any welfare gain would become increasingly small.

⁴ The semi-log schedule was used by Cagan in his classic paper (1956). Since the real rate was very small, as compared to the rate of inflation under hyperinflation, he used the expected rate of inflation as the opportunity cost of holding real balances.

⁵ This result simply states that a monopolist facing a demand curve with a constant elasticity less than unity can always increase total revenue by raising relative price.

expectations are adaptive. If inflation is held constant, expectations catch up and the revenue falls below its steady state level. If the authorities attempt to exploit this lag repeatedly, a constant coefficient on adaptive expectations would imply serially-correlated errors, or the coefficient itself would be revised. Cagan's explanation rests on the joint hypothesis of adaptive expectations and the semi-log schedule.

Lucas (1994) sees no problem in extending the double-log to hyperinflation. Indeed, it clears up Cagan's overshooting paradox. With the double log, additional revenue always accrues at higher inflation. Moreover, no lag exists since Lucas assumes expectations are rational.

A constant growth of money (above Cagan's steady state maximum) would provide a controlled experiment allowing us to distinguish between these two composite hypotheses. If Cagan's joint hypothesis is correct, the revenue should first rise and then fall permanently. If Lucas' joint hypothesis is correct, the revenue should be unchanged. The real world does not provide us with neat experiments, however. Nor would the experiment allow us to distinguish between aspects of the composite hypothesis: With either demand function, no lag occurs if expectations are rational.

The German hyperinflation lasted for only a few years. The pattern of revenue was oscillating: It rose and fell in cycles depending on the fiscal needs of the authorities. In any case, the Box-Cox transformations come down squarely on the side of the semi-log. Is the Cagan overshooting paradox explainable by a regime change so that adaptive expectations are—for a time—rational? Or, even under rational expectations and the semi-log money demand function, do we have a case of time inconsistency? Here is a fascinating area of research for which data are available.

OPTIMAL TAX ON REAL BALANCES

We next turn to how the use of the double log rather than the semi-log affects the analysis of the optimal tax on cash

(which well may be zero). Bailey used the average welfare-cost-per-dollar of revenue to measure the social costs of seignorage. Later, it was suggested that the marginal welfare-costs-per-dollar be set equal to the average of the distortions due to other taxes. General equilibrium theorists such as Ballard, Shoven and Walley (1985) have calculated the average value of these marginal welfare-costs-per-dollar of revenue for distortionary taxes. Including a tax on cash balances in the menu of taxes, the procedure for calculating the optimal tax rate is to set the marginal revenue from money creation equal to the average of the marginal distortions due to other taxes. This calculation accepts, in the spirit of the second best, the preexisting distortions caused by other taxes and assumes that any additional revenue occurring from the tax on cash is used for exhaustive expenditure rather than rebated to the private sector in the form of lower taxes. This procedure is one of many possible ways of handling second best optimal tax policy.

Given a constant real rate, the rate of price change sets the money rate of interest. For any demand function,

$$\frac{M}{P} = \phi(r),$$

the revenue is defined as the money rate (r) times the quantity of money demanded at that rate, $R = r\phi(r)$. The ratio of marginal welfare cost to the marginal increment to revenue is a function of the interest elasticity as given by

$$\frac{dW}{dr} / \frac{dR}{dr} = \frac{\eta}{1-\eta},$$

where W is the welfare loss. This result is a variant of Ramsey's inverse elasticity rule: Tax most heavily commodities in inelastic demand. It is a partial equilibrium result that implies the shifting of resources is from the taxed to the untaxed sector so that all cross elasticities are zero. In the special case that the optimal tax structure is proportional, real balances should be not taxed at all.⁶ For the constant elasticity function,

⁶ These optimal tax rules appear to have been recently rediscovered. For an earlier account see Marty (1976a, 1976b).

$$\frac{M}{P} = Ar^{-\eta},$$

the average and marginal ratios are equal. The proof is simple: the welfare loss $W = \eta Ar^{(1-\eta)}/1-\eta$. Since the revenue is

$$r \frac{M}{P}, W/R,$$

(the average ratio) = $\eta/1-\eta$ (the marginal ratio).

With a constant elasticity, if dW/dR is greater than for other taxes, real balances shouldn't be taxed at all. But what if it is less? There appears to be no equilibrium solution. The double log highlights a problem with this approach, a failing that is independent of the demand function used.

To see this, take the semi-log function, set values for the real rate and the semi-elasticity, and then solve for the optimal money rate. This determines the optimal rate of price change that equates dW/dR for real balances to the value for other taxes.

The difficulty is that the authorities don't directly control the money rate of interest, they control the monetary base. Holding other taxes constant and varying the expenditure side makes no sense. In the double log case, if dW/dR for cash were lower than the average for other taxes, expenditures would increase without limit! The proper technique is to hold expenditures constant and optimally substitute one tax for another. No shortcut avoids estimating the welfare effects of altering other taxes.

THE COMPOSITION OF THE MONEY SUPPLY

We turn to the question of what is the proper definition of the money supply. Lucas defines the money supply as M1 and runs his welfare integrals from zero to a positive money rate—a procedure that adds high-powered money and interest bearing deposits and treats this composite as non-interest bearing. In this regard, it is relevant to return to Bailey's classic article

and review his treatment of currency and deposits.

Bailey used Cagan's data to estimate the welfare cost of the inflation. He set the real rate at zero (its level is too small to play a role in hyperinflation) so that the cost of holding currency was the anticipated inflation (a proxy for the money rate). He assumed competitive banks pay interest on deposits but are subject to a sterile legal reserve requirement. A zero profit condition was imposed: The revenue from interest-bearing assets was completely dispersed in interest payments on deposits. Under these assumptions, competition would force banks to pay on deposits (D) the money rate, r , (i.e., the rate of inflation) on loans, L , times $1-z$, the reserve ratio. Zero profits are earned since $D(1-z)r = rL$. The cost of holding deposits is rz : the difference between r (the money rate on bonds) and the interest on deposits, $r(1-z)$. Since deposits are partially indexed against the inflation, Bailey assumed that at very high rates of inflation, the public uses only deposits and all currency would be held as bank reserves. Bailey's welfare integrals run from zero to rz reflecting the partial indexing of the deposit rate to inflation. Bailey then contrasts a currency-only economy with a bank-only economy. To get the same revenue and incur the same welfare cost as in a currency economy, the bank economy can inflate at a rate equal to that in the currency world multiplied by the reciprocal of the reserve ratio. The inflation rate is Π/z , but the cost of holding deposits is Πz so that the public holds the same real balances as in a currency-only world. Although the tax rate is Π/z , the tax base is zD so the tax revenue is the same as in a currency-only world. In contrast to Lucas, for Bailey there is a substantial reduction in welfare costs since the interest on deposits is partially indexed against inflation.

How satisfactory is Bailey's treatment of the different roles of currency and deposits? In the first place, there is no model of the general case where individuals hold both currency and deposits. Their distinct roles are not analyzed. Note, for example, the comparison of the currency and the banking economy implicitly assumes that the demand

curve for currency is the same as that for deposits. Moreover, the zero-profit condition ignores any marginal costs of intermediation for which the banks would have to be compensated. Nor is this all. Bailey makes the implicit assumption—which may be correct—that individuals move from currency into deposits as the percentage point difference $\pi(1-z)$ between the yield on currency and deposits rises. At high inflation, this movement is complete and all currency is held as bank reserves.

We, like Lucas, make no claim to fully analyze the role of currency and deposits. The real issue is the polar choice between two very incomplete models, however. On the one hand, we can lump currency and deposits together and assume that both pay no interest. We can then run our integrals from zero to the money rate of interest. This is Lucas' choice. On the other hand, we could assume with Bailey that all high-powered money is held as reserves and run the welfare integrals from a money rate of zero to rz , the opportunity cost of holding interest bearing deposits. Lucas' choice overstates the welfare loss when we deviate from the Friedman rule. The assumption in Bailey, that all deposits pay interest, underestimates the welfare loss. Without a theory of banking in which the distinctive roles of currency and deposits are modeled, it remains uncertain how large is the degree of overestimation or underestimation.⁷

As an alternative to these two polar positions, one might tentatively try modeling the distinct roles of currency and deposits. Making the bold assumption that currency and deposits face the same demand function (including the elasticity), it would be possible to run separate integrals: For currency, the integration would run from a zero money rate to a positive rate: For deposits, it would run from zero to rz . The general expression for the welfare loss (W) is

$$W = \int_0^r \phi(x) dx - r\phi(r) + \int_0^{rz} \phi(x) dx - rz\phi(rz).$$

With the constant elasticity schedule $M/P = Ar^{-\eta}$, the welfare loss is

$$W = \left[\frac{A\eta}{1-\eta} r^{1-\eta} \right] + \left[\frac{A\eta}{1-\eta} (rz)^{1-\eta} \right].$$

At $z = 0$, only currency would incur a welfare loss; as z approached 1, the welfare loss on deposits would increasingly approximate the loss on currency. All these models make the unsatisfactory assumption that the same demand schedule applies to currency and deposits. Moreover, the intermediate model assumes that the currency-deposit weights remain unchanged, which is unlikely during high rates of inflation.

Despite these defects, our preference—in an era of increasing financial deregulation of interest rates—is to go the intermediate route.

We came across Ed Prescott's comments on a paper by King and Wolman (1996) towards the completion of this paper. Prescott criticizes the authors' analysis:

The theory has households holding non-interest bearing money, while the monetary aggregate used in the demand for money function is M1. Most of M1 is not non-interest-bearing debt held by households. Only a third of M1 is currency and half of that is probably held abroad. Another third is demand deposits held by businesses, which often earn interest *de facto*. Households do not use these demand deposits to economize on shopping time. The final third is demand deposits held by households that, at least in recent years, can pay interest.

This criticism is also applicable to Lucas' treatment.

CONCLUSION

Our discussion has shown skepticism about Lucas' extension of the double log schedule to both very low as well as very high money rates of interest (periods of hyperinflation). Since data are not available for the American economy at very low money rates, we have discussed the implications of using the double log rather than the

⁷ Lucas is quite upfront about the need for an explicit theory of banking to distinguish between the separate roles of currency and deposits, but he leaves matters there.

more commonly used semi-log in hyperinflation. This is a fascinating area for research since data are available in hyperinflation periods.

Turning to the gains from going to very low rates of interest in the American economy, it would appear that Lucas overestimates these gains. If M1 is used as a relevant money supply, some correction should be made for the interest paid on portions of M1. (We contrasted Bailey's treatment with Lucas'.) Moreover, more than half of U.S. currency is held abroad by foreigners. Note that this affects any estimate of the welfare gains when either the monetary base or M1 is used. What remains is a cautionary tale. We should treat any estimates of the welfare costs with caution when the distinct roles of currency and deposits are not modeled.

Mundell, Robert A. "Inflation and Real Interest," *Journal of Political Economy* (June 1963), pp. 280-83.

Phelps, Edmund S. "Inflation in the Theory of Public Finance," *Swedish Journal of Economics* (March 1973), pp. 67-82.

Ramsey, Frank P. "A Contribution to the Theory of Taxation," *Economic Journal* (1927), pp. 47-61.

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

Tobin, James. "Money and Economic Growth," *Econometrica* (October 1965), pp. 671-84.

REFERENCES

Auernheimer, Leonardo. "The Honest Government's Guide to the Revenue from the Creation of Money," *Journal of Political Economy* (May/June 1974), pp. 598-606.

Bailey, Martin J. "The Welfare Cost of Inflationary Finance," *Journal of Political Economy* (April 1956), pp. 93-110.

Bali, Turan. "Essays in Macroeconomics and Finance," unpublished Ph.D. dissertation, City University of New York, 1998.

Ballard, Charles L., John B. Shoven, and John Whalley. "General Equilibrium Computations of the Marginal Welfare Costs of Taxes in the United States," *The American Economic Review* (March 1985), pp. 128-38.

Cagan, Phillip. "The Monetary Dynamics of Hyperinflation," *Studies in the Quantity Theory of Money*, M. Friedman, ed., University of Chicago Press, 1956, pp. 25-117.

Friedman, Milton. *The Optimum Quantity of Money and Other Essays*, University of Chicago Press (1969).

King, Robert G., Alexander L. Wolman. "Inflation Targeting in a St. Louis Model of the 21st Century," this *Review* (May/June 1996), pp. 83-107.

Lucas, Robert E., Jr. "On the Welfare Cost of Inflation," manuscript, Department of Economics, University of Chicago, 1994.

Marty, Alvin L. "A Note on the Welfare Cost of Money Creation," *Journal of Monetary Economics* (January 1976a), pp. 121-24.

_____. "Real Cash Balance and the Optimal Tax Structure," in *Essays in Economic Analysis, Proceedings of the Association of University Teachers of Economics Annual Conference, Sheffield 1975*, Michael J. Artis and A. Robert Nobay, eds., Cambridge University Press, 1976b.