Economic Quarterly—Volume 99, Number 2—Second Quarter 2013—Pages 83–116

The Monetarist-Keynesian Debate and the Phillips Curve: Lessons from the Great Inflation

Robert L. Hetzel

A chievement of consensus over the cause of cyclical fluctuations in the economy and the nature of inflation has foundered on the impossibility of running the controlled experiments that isolate a single cause from the multiple forces that impact the economy. In this respect, the period from the mid-1960s through the end of the 1970s (the Great Inflation) is important in that the characterization of monetary policy—the economists' proxy for an experiment—was unusually clear.[1] Monetary policy was activist in that the Federal Reserve pursued both unemployment and inflation objectives in a way shaped by the assumed tradeoffs of the Phillips curve.[2] The experience of the Great Inflation did produce enduring changes, especially the assumption of responsibility by central banks for the control of inflation without recourse to wage and price controls. However, the

The author gratefully acknowledges helpful comments from Thomas Lubik, Andrew Owen, Felipe Schwartzman, and Alex Wolman. The views expressed in this article are those of the author and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: robert.hetzel@rich.frb.org.

[1] Much of the commentary in this article summarizes work by Hetzel (1998; 2008a, Chs. 5–12, 22–25; 2012, Ch. 8; and 2013a).

[2] Over time, economists who urge an activist policy aimed at achieving an optimal mix of low inflation and low unemployment or an optimal tradeoff in the variability of these variables have altered the character of the empirical correlations between inflation and unemployment to which they attribute structural significance. Until the end of the 1970s, the period relevant for the discussion here, most commonly, they emphasized the correlation between inflation and the unemployment rate. Subsequently, they have emphasized the correlation between the difference in the unemployment rate and a reference value often termed the NAIRU (non-accelerating inflation rate of unemployment) and the change in the rate of inflation.

difficulty of isolating the impact of policy from other forces, especially inflation shocks, has limited the conclusions that economists draw from this experience.

In the 1960s, and well into the 1970s, an unusual degree of professional consensus existed. This Keynesian consensus emerged out of two dramatically contrasting episodes. The persistence of high unemployment in the decade of the 1930s (the Great Depression) appeared to demonstrate the weak equilibrating properties of the price system. In contrast, the low unemployment during World War II appeared to demonstrate the usefulness of fiscal policy in managing aggregate demand in order to maintain employment at its full employment level.

Supported by this intellectual consensus during the Great Inflation, policy attempted to stabilize unemployment at a lower level than had prevailed over most of the post-War era. The activist policy pursued in order to achieve this objective engendered the monetarist-Keynesian debate, which centered on whether policymakers could and should base policy on the observed inflation-unemployment relationship captured by the empirical correlations of the Phillips curve.

Section 1 offers a broad overview of the methodology economists use for learning from historical experience—whose antecedents lie in the Friedman-Cowles Commission debate of the early 1950s. Section 2 summarizes the way in which the contemporaneous understanding of the Phillips curve shaped monetary policy in the 1970s. Sections 3 and 4, respectively, contrast Keynesian and monetarist views on the Phillips curve and the resulting disagreement over the desirability of an activist monetary policy. Section 5 explains the way in which the Samuelson-Solow interpretation of the Phillips curve embodying an inverse relationship between inflation and unemployment supported the policy of aggregate-demand management in the Great Inflation. Section 6 reviews the challenge made by Milton Friedman to the Samuelson-Solow interpretation of the Phillips curve. In a way analogous to the contrasting experiences of the Great Depression and World War II, Sections 7 and 8 summarize how the contrasting experiences of the Great Inflation and the Volcker-Greenspan era changed the prevailing Keynesian intellectual consensus. The article concludes, in Sections 9 and 10, with some speculation on the course of the current debate over the causes of the Great Recession, which began in earnest in 2008.

1. FRIEDMAN AND THE COWLES COMMISSION ECONOMISTS: COMPLEMENTARY ADVERSARIES

In the late 1940s, the University of Chicago and the University of Cambridge assembled perhaps the greatest collection of intellectual brilliance the economics profession will ever see. They provided much of the impetus involved in changing economics from its then dominant institutionalist character to the neoclassical character now considered mainstream. Along with the mathematical formalization of Keynes's (1936) book (*The General Theory of Employment, Interest and Money*), in Hicks (1937) the methodology developed by the economists of the Cowles Commission laid out the general framework for construction of models of the economy and highlighted the econometric issues of identification of structural equations from the reduced-form correlations found in the data.[3] In his essay "The Methodology of Positive Economics," Friedman ([1953a] 1953) criticized the identification strategy of the Cowles Commission with its reliance on a priori assumptions about which variables could be excluded in the estimation of the equations comprising a model of the economy.[4]

Friedman argued that many alternative models would fit a set of macroeconomic time series equally well.[5] As a consequence, goodness of fit for a given body of data would not distinguish between models. Hypothesis testing requires the elucidation of contrasting implications of alternative models. Those contrasting implications then should be taken to data sets not available to the economist at the time of building the model. Most notably, testing required that models not only fit the existing data but also that they yield implications about the future.[6]

Understanding the context of Friedman's 1953 essay helps to elucidate the statements it contains about hypothesis testing. At the end of the 1940s, there was an effort to test the marginal foundation of neoclassical economics by examining its "realism," for example, through surveys asking the managers of firms whether they choose price and

[3] The Cowles Commission pioneered the representation of the economy by a system of stochastic difference equations. As expressed by Tjalling Koopmans (1947, 167), the Cowles Commission's members worked on empirical estimation based on recognition of the fact that "the mere observation of regularities in the interrelations of variables ... does not permit us to recognize or to identify behavior equations among such regularities." The general approach of giving the behavioral equations that represent the economy a microeconomic foundation shapes the research agenda of macroeconomics.

[4] Sims (1980) talked about "incredible" identifying restrictions of the large-scale econometrics models spawned by the Keynesian attempt to give empirical content to the Cowles Commission agenda.

[5] See Chari, Kehoe, and McGrattan (2009) for a similar statement.

[6] For a restatement, see Friedman and Schwartz (1991).

output based on a marginal cost schedule. The then-dominant institutionalist school questioned the realism of marginal cost pricing. Friedman argued that the theoretical assumptions of neoclassical models were a necessary abstraction required in order to yield refutable implications.[7] The relevant test of a model is its predictive ability. Because of its complexity, a "realistic" model would always afford a rationalization of the data but the economist could not distinguish between fitting a model to the data and testing its validity.

Beyond the simplification entailed by the theoretical abstraction necessary to compare the implications of a model to the data in a way capable of refuting rather than rationalizing the model, it is necessary to separate exogenous from endogenous variables. The ideal is the controlled experiment of the physical sciences. A test of the competing hypotheses that guide the formulation of alternative models is then simplified because of the assignment of causality made possible by the controlled experiment. Applied to economics, the Friedman strategy was to relate both the evolution of central bank procedures and episodes of significant departures from those procedures to changes in the political and intellectual environment unrelated to the operation of the price system. This diversity of central bank behavior serves as a semi-controlled experiment informative for disentangling causation in the historical association between real and monetary instability.

The spirit of the Friedman approach to testing models involves, as a first step, specification of the alternatives. At this stage, models can be superior along two dimensions. First, some may be better microfounded than others. Second, some may explain a more challenging set of empirical phenomena. That is, they are more resistant to fitting time series through data mining. The ideal is to proceed along two parallel, inter-related paths: model building and the isolation of "robust" correlations.

The search for robust correlations requires searching across time and across countries in pursuit of persistent relationships. In the context of monetary models of the business cycle, correlations between monetary and real instability that survive this diversity of experience are as close as one can come to a controlled experiment. The diversity of experience limits the possibility of some nonmonetary cause common to all episodes producing the correlation between monetary and real instability. The discipline of looking at the entire set of historical experiences rather than isolating individual episodes favorable to one hypothesis, in this case, the monetary nonneutrality explanation

[7] Of course, they also impose the discipline of constrained optimization that households and firms undertake all available trades that improve their welfare (markets clear).

of the business cycle, reveals whether real instability arises in contexts of monetary stability as well as in contexts of extreme monetary instability.

Specifically, the economist looks for event studies, that is, episodes in which he (she) has some information particular to the time period about the nature of causation. Because of the impossibility of controlling for extraneous forces in particular episodes, the ideal is one where metastudies generalize across a wide variety of historical event studies. In particular, do monetary-real correlations appear in a sufficiently wide variety of historical episodes so that the only common element in the episodes is likely to be the behavior of the central bank? Correlations that persist across time and place and come tagged with information of central bank behavior unrelated to the stabilizing operation of the price system then become the "stylized facts" that discipline the choice of frictions to incorporate into models.[8]

The challenge is to run a horse race among models that potentially selects the one that is likely to offer better predictions out-of-sample. Although alternative models can differ in the adequacy of their microfoundations, the Friedman emphasis is on the assumption that each model builder knows the data and will select a combination of model and data that support his (her) model. By itself, neither model fit nor economic theory is adequate to identify the true structural equations. One central element in model selection is to discipline the horse race through identification of policy using a variety of historical information rather than representing policy by a general functional form with free parameters the estimation of which will necessarily aid the fit of any model.

To make the discussion more specific, a correlation common to all recessions is central bank behavior that imparts inertia to reductions in interest rates while the economy weakens. For central banks concerned with the behavior of the external value of their currency, this behavior is associated with countries going onto the gold standard or a peg with a foreign currency at a parity that overvalues the domestic currency (requires a reduction of the real terms of trade through deflation). For the other cases, this behavior is associated with a concern to lower inflation or asset prices considered artificially elevated by speculation. These episodes come tagged with information that the behavior of the central bank does not arise out of a systematic reaction function related

[8] One problem in macroeconomics is the practical difficulty of generalizing from the vast literature on historical episodes that are potentially useful as event studies. This difficulty makes it harder to reach agreement in monetary economics over the "stylized" facts a model should explain. In contrast, new mathematical techniques useful in model construction are more readily incorporated into mainstream models.

to the ongoing behavior of the economy. Monetarists point to such a correlation as robust.

In monetary economics, the horses in these races divide into three basic classes. In the Keynesian tradition, cyclical fluctuations arise from real shocks in the form of discrete shifts in the degree of investor optimism and pessimism about the future large enough to overwhelm the stabilizing properties of the price system and, by extension, to overwhelm the monetary stimulus presumed evidenced by cyclically low interest rates. In the quantity theory tradition, cyclical fluctuations arise from central bank behavior that frustrates the working of the price system through monetary shocks that require changes in individual relative prices to reach, on average, a new price level in a way uncoordinated by a common set of expectations. In the real-business-cycle tradition, cyclical fluctuations arise from productivity shocks passed on to the real economy through a well-functioning price system devoid of monetary nonneutralities and nominal price stickiness. Of course, only the first two horses contended in the debate during the Great Inflation.

2. THE CENTRAL ROLE OF THE PHILLIPS CURVE DURING THE GREAT INFLATION

The Phillips curve is a set of empirical observations showing an inverse relationship between the behavior of inflation and unemployment. At the heart of the activist policy pursued during the Great Inflation was the belief in an "exploitable" Phillips curve, that is, a Phillips curve allowing the policymaker to trade off between the achievement of unemployment and inflation objectives. The monetarist-Keynesian debate turned, to a significant extent, on the issue of whether the empirical correlations of the Phillips curve represented a structural relationship that would allow policymakers to trade off between their pursuit of the two variables, with predictable consequences.[9]

Specifically, during periods of economic recovery from a cyclical trough when inflation had fallen and the unemployment rate was above normal and thus unemployment had become the main concern, policymakers assumed that monetary policy could be expansionary without

[9] During the Great Inflation, monetary policymakers eschewed the language of tradeoffs. As a result, discussions within the Federal Open Market Committee (FOMC) never explicitly employed the conceptual framework of the Phillips curve. Moreover, FOMC discussion followed the packaging for the public of policy actions as individual actions, each of which was defensible in a common sense way in the context of the contemporaneous behavior of the economy and the resulting relative priority assigned to achieving unemployment and inflation objectives. As a result, both the systematic character of monetary policy and the conceptual framework generating that policy have to be inferred by economists.

exacerbating inflation. That is, a flat Phillips curve would allow a reduction in unemployment to its full employment level with little increase in inflation. In the aftermath, in the advanced stages of economic recovery when a reduction in unemployment and an increase in inflation turned inflation into the main concern, policymakers assumed that monetary policy could be restrictive by creating a moderate, socially acceptable increase in unemployment. That is, a moderate but sustained increase in unemployment above its full employment level acting through a downward-sloping Phillips curve would lower inflation at an acceptable social cost in terms of unemployment. In a way given by the sacrifice ratio embedded in the Phillips curve, monetary policy could engineer the required number of man-years of excess unemployment—the so-called soft landing—through an extended but moderate increase in unemployment above its full employment level.

This common understanding of the nature of the Phillips curve and activist policy rested on two basic assumptions. First, inflation is a nonmonetary phenomenon. That is, inflation springs from a variety of real factors rather than from the failure of the central bank to control money creation. One reason that the Great Inflation is an interesting laboratory for economists was the existence of a monetary aggregate (M1) that provided a good measure of the stance (stimulative or contractionary) of monetary policy due to the interest-insensitive nature of real money demand and a stable, albeit lagged, relationship with nominal expenditure. However, the assumption that money responded passively to the various real forces that determine the combined total of real aggregate expenditure and inflation (nominal aggregate expenditure) removed money from consideration as a useful policy instrument. It was the real character of inflation that made the Phillips curve, rather than money, into the relevant predictor of inflation.

The second basic assumption was that policymakers understood the structure of the real economy sufficiently well to pursue an unemployment objective. They knew the level of unemployment consistent with full employment, by consensus, taken to be 4 percent. The excess of unemployment over this full employment level measured the amount of idle workers desiring productive employment. Also, policymakers could forecast the behavior of the economy based on their choice of policy sufficiently well to exploit the tradeoffs of the Phillips curve. They could lower excess unemployment through stimulative monetary policy at an acceptable cost in terms of inflation. Analogously, when the unemployment rate became an intermediate objective of policy central for lowering inflation rather than an objective in itself and policy was restrictive, they could manage inflation with an acceptable cost measured in terms of extended excess unemployment.

3. AN OVERVIEW OF TRADITIONAL KEYNESIAN VIEWS

As described in *The General Theory*, swings in investor sentiment, which Keynes termed animal spirits, drove the business cycle. Adjustment to these swings in sentiment occurred through changes in output unmitigated by the operation of the price system. Keynes fixed nominal prices by assuming rigid wage rates and by taking the price level as an institutional datum. The resulting framework served as a clarion call for government action to counter recession. It did so by challenging the prevailing view that the deflation and recession following the bursting of an asset bubble required an extended period of rectifying accumulated imbalances (Hetzel 1985; 2012, Ch. 3).

In Keynes's framework, the exogeneity of fluctuations in investment captured the assumption that irrational swings from optimism to pessimism about the future overwhelm the ability of the stabilizing properties of the price system. That is, in recession, no decline in the real interest rate is sufficient in order to redistribute demand from the future to the present to maintain aggregate demand equal to potential output. In response to an exogenous decline in investment, output has to decline. Otherwise, given the exogenous decline in investment, the full employment level of saving would exceed investment. A decline in output is necessary to reduce saving in line with a lower level of investment.

However, a given decline in output decreases saving by only a fractional amount because of a marginal propensity to consume out of income (output) greater than zero. The required reduction in saving must occur through a decline in output (income) that is a multiple of the decline in investment. As captured by the Keynesian multiplier, exogenous swings in investment translate into shifts in output in a mechanical way based on the inverse of the marginal propensity to save (one minus the marginal propensity to consume). The optimism in Keynes's message came from the implication that the government could offset the excessive private saving that arose at full employment through public dissaving, that is, through deficit spending. With social saving (government dissaving plus private saving) at the full employment level, output need not fall in order to equate private saving to a lower level of exogenous investment.

At a deeper level, the issue is why an increased desire to save (transfer resources to the future) in order to guard against a future that has become darker and more uncertain does not translate into increased investment but instead requires a decline in output. That desire is frustrated on two levels. The ability of financial intermediation to transfer resources from savers to investors with opportunities for

productive investment breaks down.[10] Also, the nominal rigidity of wages and prices frustrates the desire to save for the future through an increased work effort. Without the management of aggregate demand by government through deficit spending, output and employment can fall short of potential output over extended, perhaps indefinite, periods.

Keynesians believed that the central bank should target the behavior of the unemployment rate (the amount of idle resources in the economy due to the weak ability of the price system to maintain full employment and the full utilization of resources). The central bank should pursue this real objective subject to the constraint imposed by the acceptable level of inflation. The central role of the Phillips curve derived from the assumption that it offered policymakers a practical way of estimating the cost in terms of inflation incurred by the pursuit of the full employment objective. Similarly, in response to inflation shocks, the Phillips curve allowed policymakers to predict the cost in terms of excess unemployment of mitigating the inflation produced by the inflation shock.

4. AN OVERVIEW OF MONETARIST (QUANTITY THEORY) VIEWS

Monetarism, as formulated by Milton Friedman, challenged the activist monetary policy pursued during the Great Inflation and the Keynesian consensus that supported it. Monetarists believed that the central bank should concentrate on the control of money creation with the objective of price stability. This monetary objective would turn over to the price system the exclusive responsibility for the determination of real variables like the unemployment rate.[11] The following elucidates the central role played by the need for monetary control.

Although central banks use the interest rate as their instrument, their uniqueness comes from monopoly control over the monetary base (bank reserves and currency). Because the monetary base is the medium used to effect finality of payment in transactions for whatever instruments possess the property of a medium of exchange (broad money or simply money here), the control of money creation requires the control of the monetary base. It follows that the interest rate rule the central

[10] A liquidity trap (the willingness of the public to hold whatever amount of money the central bank creates) vitiates the effectiveness of monetary policy as opposed to fiscal policy.

[11] The intensity shown by Keynesians in the monetarist-Keynesian debate came from the fear that a central bank policy organized around monetary control would lead to a rule for controlling money that left the determination of real variables to the operation of the price system.

bank follows must provide for that control. The following elucidates the discipline imposed on that rule.

Money serves three functions. It is a numeraire, a store of value, and a medium of exchange. In order to serve its function as a numeraire, the money price of goods (the number of dollars that exchange for a representative basket of goods consumed by households) must evolve predictably. The simplest case is that of price stability. In its function as a numeraire, money has a public good aspect. Although firms set prices in terms of dollars, they only intend to set a relative price (the rate of exchange of their product with other products). There is then an advantage to all firms that set dollar prices for multiple periods in setting the dollar price for their product based on the same assumption about the future price level. An assumption of rational expectations is that the central bank can organize this coordination by following a rule that causes the price level to evolve predictably.[12] In the sense of Hayek (1945), a stable numeraire is one element in allowing the price system to economize on the information that households and firms need in order to make decisions.

Money also serves as a medium of exchange. To effect transactions, the public desires to hold a well-defined amount of purchasing power (the nominal quantity of money multiplied by the goods price of money, the inverse of the price level). To prevent an unpredictable evolution of the price level that vitiates the role of money as a numeraire, the central bank must cause nominal money to grow in line with the real demand for money consistent with growth in potential output plus transitory demands. Even if central banks do not have money targets and even if money does not serve to forecast economic activity, monetary stability requires that central bank procedures control money creation.[13]

A monetary-control characterization of policy follows if the price level is a monetary phenomenon in the strong form in the sense that there is no structural (predictable) relationship between real variables like unemployment and nominal variables like nominal money and the monetary base, the variable over which the central bank exercises

[12] The assumption is not true in any literal sense in that the evolution of the monetary standard since the breakdown of the gold standard has been one of learning. However, it possesses the powerful implication that if the central bank behaves in a credible, consistent way, its rule will discipline the way in which markets forecast inflation.

[13] Like any abstraction, one has to give empirical content to the variable "money." In principle, one would like a measure of the transactions (liquidity) services yielded by different assets, such as contained in a Divisia aggregate (Barnett 1982). A complicating factor is that, since 1994, the Federal Reserve Board has not measured the extent to which banks "sweep" deposits off their balance sheets in order to avoid the tax imposed by non-interest-bearing reserve requirements. Monetary aggregates like M1 are therefore likely mismeasured.

ultimate control. Two implications follow from the absence of a structural relationship between money and real variables. First, the central bank must provide a nominal anchor. Because the welfare of individuals depends on real variables (physical quantities and relative prices), nothing in their behavior gives money a well-defined value in exchange for goods by limiting its quantity. The intrinsic worthlessness of money requires the central bank to follow a rule that limits the nominal quantity of money.

The second implication of the absence of a structural relationship between money and real variables is that in order to provide for monetary and real stability, the central bank must turn over the determination of real variables to market forces. In this sense, in order to provide for monetary stability, the central bank must avoid "price fixing" by interfering with the operation of the price system. Equivalently, given that central bankers use an interest rate as their policy instrument, in order to provide for monetary and real stability, monetary policy procedures must entail moving the nominal interest rate so that the resulting real interest rate tracks the natural interest rate.[14] Specifically, central banks must allow market forces to determine the real interest rate and, by extension, other real variables like the unemployment rate.[15]

The control of trend inflation then comes from the way in which the central bank's rule creates a stable nominal expectational environment that shapes the way in which firms in the "sticky" price sector set prices for multiple periods rather than through manipulation of an output gap based on Phillips curve tradeoffs. A critical facet of the monetarist assumption that the price system works well in the absence of monetary disorder is rational expectations.[16] Specifically, when firms set a dollar price for their product for multiple periods, they take into account the way in which future changes in the price level will affect the relative price of their product. The assumption of rational expectations implies that if the central bank behaves in a predictable and credible way, firms collectively will coordinate these relative-price maintaining changes in

[14] In the context of the New Keynesian model, the natural rate is the real interest rate that would obtain in the absence of any nominal rigidity in prices. The counterpart in the writings of Milton Friedman is the assumption that the price system gives real variables well-defined (natural) values when actual and expected inflation are equal.

[15] This Wicksellian view contrasts with the Keynesian view in which multiple sources of price stickiness exist, say, in the setting of wages and product prices. In principle, if the central bank possessed sufficient knowledge of the economy, it could follow a rule that managed real aggregate demand by controlling the real interest rate in order to trade off optimally between inflation and both employment and output gaps. See the Appendix.

[16] This assumption is not in Milton Friedman's formulation of the quantity theory. It first appears in the mathematical formulation of monetarist ideas in Lucas ([1972] 1981).

dollar prices on the central bank's inflation target. The self-interest of firms in setting their markup of price over marginal cost optimally over time causes them to use information efficiently about the nature of the monetary regime.

Individually, firms set relative prices based on marginal cost. The central bank's rule separates the determination of the price level from the determination of relative prices (at cyclical and lower frequencies). As a consequence of following a rule that causes the real interest rate to track the natural interest rate (the real rate determined by market forces), the central bank allows the price system to determine real variables and allows the price system to keep real output fluctuating around its potential level.[17] As a consequence of its interest rate target, the central bank then allows nominal money to grow over time in line with the real money demand associated with growth in potential output. The interest rate target also allows changes in money to accommodate transitory changes in money demand and whatever inflation occurs as a consequence of the central bank's inflation target. In this way, the rule causes nominal money to grow over time in a way that does not require unanticipated changes in the price level in order to bring real money into line with real money demand.

The central bank can control trend inflation—no less and (just as important) no more. In order to avoid destabilizing economic activity, it should allow transitory noise to pass through into the price level. In the passage containing the famous "long and variable lags" phrase, Friedman (1960, 86–8) argued that the power of the central bank was limited to the ability to control trend inflation. Any attempt to manage the behavior of the real economy or to smooth transitory fluctuations in inflation would in practice destabilize the economy due to policymakers' lack of knowledge of the structure of the economy. The following summarizes the experiment with aggregate demand management in the decade and a half after mid-1965.[18]

5. THE VAST EXPERIMENT OF PAUL SAMUELSON AND ROBERT SOLOW

In *The General Theory*, Keynes assumed that with excess capacity in the economy increases in aggregate demand would raise output. Only

[17] As noted above, Keynesians point to the low rates of interest in recession as evidence of the impotence of monetary policy. Monetarists point to the inertia central banks put into the interest rate when the economy weakens and the associated monetary deceleration. A low interest rate in recession implies only that the public is pessimistic about the future.

[18] For other accounts, see Hetzel (2008a, 2013a) and King (2008).

at full employment would increases in aggregate demand appear as price rises.[19] Given the general consensus that emerged after World War II that a 4 percent or lower unemployment rate represented full employment, an unemployment rate above 4 percent implied the existence of idle workers—workers who wanted to work at the prevailing wage rate but could not find work. Aggregate demand management should then be able to push the unemployment rate down at least to 4 percent without inflation. In the language of the time, demand-pull inflation would not be a problem.

The contest for the presidency between John F. Kennedy and Richard Nixon in 1960 initiated a national debate over the use of aggregate-demand management to lower the unemployment rate to 4 percent or lower. Kennedy's economic advisers wanted to pursue an activist policy of aggregate demand management. Politically, the chief obstacle to adoption of such a policy with its deliberate deficits was fear of inflation. The Kennedy Council of Economic Advisers needed a model that would predict the inflation rate associated with the reduced unemployment rate presumed to follow from a policy of aggregate-demand management. The Samuelson-Solow ([1960] 1966) interpretation of the empirical correlations of the Phillips curve provided those predictions.

Consistent with the Keynesian temper of the time, Paul Samuelson and Robert Solow offered an interpretation of the Phillips curve based on the premise that inflation is a real phenomenon rather than a monetary phenomenon. As a real phenomenon, there is no single explanation for inflation. The Keynesian taxonomy of the causes of inflation contained two kingdoms. Aggregate-demand (demand-pull) inflation arises from a high level of aggregate demand that stresses the rate of resource utilization. Cost-push inflation arises from increases in relative prices particular to individual markets that pass through permanently to the price level. A wage-price spiral could turn cost-push inflation into sustained inflation.

For the years 1861 to 1957 for Great Britain, A. W. Phillips (1958) demonstrated the existence of an inverse relationship between the rate of change of money wages and the unemployment rate. In 1960, Samuelson and Solow ([1960] 1966, 1,347) presented a graph of the same variables for the United States. Collectively, the observations in the Samuelson-Solow graph did not exhibit any particular pattern. The two economists argued, however, that the inverse relationship found by

[19] See Keynes ([1936] 1973, 300–1). He referred to the inflation that would arise as the economy approached full employment as "bottleneck" inflation. Before full employment, cost-push inflation could occur caused by "the psychology of workers and by the policies of employers and trade unions."

Phillips appeared in two periods: 1900–30 (omitting World War I), and 1946–58. The Phillips curve had, however, shifted up in the latter period.[20]

Samuelson and Solow ([1960] 1966, 1,348) assumed that the empirical Phillips curve they identified was "a reversible supply curve for labor along which an aggregate demand curve slides.... [M]ovements along the curve might be dubbed standard demand-pull, and shifts of the curve might represent the institutional changes on which cost-push theories rest." They believed that the Phillips curve offered an exploitable tradeoff. Breit and Ransom (1982, 128) quoted Solow:

> I remember that Paul Samuelson asked me when we were looking at the diagrams for the first time, "Does that look like a reversible relationship to you?" What he meant was, "Do you really think the economy can move back and forth along a curve like that?" And I answered, "Yeah, I'm inclined to believe it," and Paul said, "Me too."

The upward shift in the post-World War II period in the empirical Phillips curve, however, created a conundrum for Samuelson and Solow over what unemployment rate to recommend as a national objective. Their graphical analysis indicated that the unemployment rate consistent with price stability (zero inflation) was 5.5 percent. That unemployment rate was unacceptable to them. Samuelson and Solow ([1960] 1966, 1,351) referred to a 3 percent unemployment rate as a "nonperfectionist's goal" and adopted it as their reference point for full employment.

The issue of what inflation rate would arise if aggregate-demand management lowered the unemployment rate to 3 percent then depended on whether the Phillips curve had shifted upward because of cost-push inflation. If not, then price stability would require an unemployment rate of 5.5 percent. Because the data did not themselves reveal whether the market power of large corporations and unions had pushed up the empirical Phillips curve of the 1950s, Samuelson and Solow ([1960] 1966, 1,350) concluded that only the "vast experiment" of targeting 3 percent unemployment could determine whether their empirically estimated Phillips curve had been pushed up by cost-push inflation. With the objective of 3 percent unemployment achieved with aggregate-demand management, in the absence of cost-push inflation, prices should be stable. If cost-push inflation did arise, government

[20] Samuelson and Solow ([1960] 1966) translated the Phillips curve of Phillips (1958) into the more familiar Phillips curve with inflation on the vertical axis by lowering nominal wage growth by an assumed rate of growth of labor productivity.

programs to deal with the market power of large corporations and unions could make price stability with full employment possible.

Samuelson and Solow ([1960] 1966, 1,347 and 1,352) accepted the possibility that an increase in inflationary expectations could have caused what they conjectured to be cost-push inflation. However, they assumed that a policy to reverse that increase in inflationary expectations would likely entail a prolonged, socially unacceptable period of high unemployment.

> The apparent shift in our Phillips curve might be attributed by some economists to the new market power of trade-unions. Thus, it is conceivable that after they [policymakers] had produced a low-pressure economy [an economy with price stability], the believers in demand-pull might be disappointed in the short run; i.e., prices might continue to rise even though unemployment was considerable. Nevertheless, it might be that the low-pressure demand would so act upon wage and other expectations as to shift the curve downward in the longer run—so that *over a decade*, the economy might enjoy higher employment with price stability than our present-day estimate would indicate. [italics added]

Samuelson and Solow warned of the social cost of maintaining the 5.5 percent unemployment rate necessary to deliver price stability if indeed inflation was of the cost-push variety. Samuelson and Solow ([1960] 1966, 1,352 and 1,353) wrote that such a "low-pressure economy might build up within itself over the years larger and larger amounts of structural unemployment" leading to "class warfare and social conflict." "[D]irect wage and price controls" were a way "to lessen the degree of disharmony between full employment and price stability."

What happened to make a reality the "vast experiment" envisaged by Samuelson and Solow? In the Eisenhower administration, the Keynesian policy prescription of aggregate-demand management exercised no practical influence because of concern for balanced budgets and for the balance of payments and gold outflows. In the 1962 *Economic Report of the President*, President Kennedy did set 4 percent as a national goal for the unemployment rate accompanied by wage "guideposts" in order to control cost-push inflation (Hetzel 2008a, Ch. 6). However, in the context of the Bretton Woods system, Kennedy was unwilling to risk a dollar crisis (a run on the dollar) given the international tension associated with the Cuban missile crisis and the Berlin Wall (Hetzel 2008a, Ch. 7). For that reason, policy remained dominated by the conservative Treasury.

Starting with the 1964 tax cut, enacted in the Johnson administration following the fall 1963 assassination of Kennedy, the political

temper turned activist. President Johnson, with roots in the tradition of Texas populism, simply disliked "high" interest rates. More important, the country split in response to the Vietnam War and the emergence of a militant civil rights movement. "Low" unemployment offered a social balm. At the same time, Keynesian economists proffered the promise of full employment, taken to be 4 percent unemployment, at an acceptable cost in terms of inflation. That promise came from a Keynesian interpretation of the Phillips curve.

With the 1964 tax cut, the political system became hostile to increases in interest rates. Congressmen argued that any such increases would thwart the will of the political system to lower the unemployment rate as evidenced by the tax cut. William McChesney Martin, chairman of the FOMC, also had to deal with an increasingly Keynesian Board of Governors. In response, he worked with Treasury Secretary Henry H. Fowler to get an income tax surcharge that would eliminate the deficit and, hopefully, remove the need for increases in interest rates. However, the temporizing that effort entailed in raising interest rates in response to strong economic growth and declining unemployment caused money growth to surge. By the end of the 1960s, 6 percent inflation had replaced the price stability (1 percent consumer price index [CPI] inflation) of the start of the decade (Hetzel 2008a, Ch. 7).

Arthur Burns replaced William McChesney Martin as chairman of the FOMC in February 1970. Burns was willing to implement an expansionary monetary policy under the condition that President Nixon would impose wage controls in order to control inflation (Hetzel 1998, 2008a). Burns got those controls in August 1971. The United States also got the "vast experiment" envisaged by Samuelson and Solow: a policy of aggregate demand management intended to create a low unemployment rate accompanied by price controls to restrain cost-push inflation.

Over time, the Phillips curve that Samuelson and Solow identified for the United States shifted. Stockman (1996, 906 and 904) shows the Phillips curve for consecutive time periods. After a noisy start from 1950 to 1959, the curve exhibited a negative slope in the 1960s. It then shifted up from 1970 to 1973 and then again in 1974 to 1983. The curve shifted down after 1986. Initially, both Keynesian economists and policymakers interpreted the upward shift in the 1970s as evidence of cost-push inflation.

6. AN EXPECTATIONS-ADJUSTED PHILLIPS CURVE: FRIEDMAN'S CHALLENGE TO SAMUELSON-SOLOW

In their challenge to the Keynesian consensus in favor of an activist monetary policy, Friedman and Schwartz (1963a) organized the data on money and the business cycle using the National Bureau of Economic Research methodology of leading, coincident, and lagging indicators. The historical narrative in Friedman and Schwartz (1963b) associated changes in the behavior of money (changes in a step function fitted to money growth rates) to behavior of the central bank adventitious to the working of the price system. This procedure isolated changes in nominal money arising independently of changes in real money demand. Friedman then used these temporal relationships to forecast both the cyclical behavior of the economy and the rising inflation during the Great Inflation.

Friedman and Meiselman (1963) also published an article showing that money, but not investment, predicted nominal output. The Keynesian assumption was that velocity would adjust in order to make whatever amount of money existed compatible with a level of nominal output independently determined by real forces. This variability in velocity should have limited the predictive power of money. The response by Ando and Modigliani (1965) provided an impetus to the construction of large-scale macroeconomic models as a way of measuring the impact of changes in investment based on structural relationships rather than the reduced-form relationships of Friedman and Meiselman. Keynesians believed that such models would allow forecasts of the evolution of the economy under alternative policies. The intention was to enable an activist policy to improve on the working of the price system, which the Keynesian consensus assumed worked only poorly to maintain the full employment of resources.

Friedman challenged the feasibility of such models. Friedman (1960) argued that "long and variable lags" inherent in the impact of discretionary policy actions could destabilize the economy. In his presidential address to the American Economic Association, Friedman ([1968] 1969) argued that economists lacked the knowledge required to construct proxies for resource slack (underutilization of resources). The large-scale econometric models required to implement an activist monetary policy necessitated measures of these output gaps. Moreover, any attempt to use monetary policy to control the behavior of a real variable like unemployment in a systematic, predictable way would cause the assumed structural equations of these models to change in unpredictable ways.

Specifically, Friedman ([1968] 1969) criticized the idea of an exploitable Phillips curve tradeoff between inflation and unemployment.[21] Friedman's criticism reiterated his belief in the monetary rather than the real nature of inflation. The correlation between nominal and real variables at cyclical frequencies arises from monetary nonneutrality due to monetary disturbances.[22] Any systematic attempt by the central bank to lower unemployment through inflation would founder on the effort of the public to forecast inflation in order to set relative prices optimally. The Phillips curve would then be vertical. This proposition came to be known as the natural rate hypothesis.[23]

This formulation of the natural rate hypothesis derived its predictive content from the distinction between anticipated and unanticipated changes in inflation. Friedman expressed that distinction in the "expectations-adjusted" Phillips curve. That is, variation in the unemployment rate is related not to variation in the inflation rate, but to variation in the inflation rate relative to the inflation rate expected by the public. Surprise changes in inflation can cause actual and expected prices to diverge and thus affect real variables. The short-run nonneutrality of money then corresponded to the interval of time required for the public to adjust its expectations in response to a higher inflation rate.

Friedman predicted that an attempt by the Fed to peg the unemployment rate at a level less than the natural rate (the value consistent with equality between actual and expected inflation) would require increased inflation. He argued that the level of the Phillips curve would shift upward as the public's expectation of inflation rose (see Humphrey [1986]). Friedman also assumed that the public formed its expectation of inflation based on the past behavior of inflation (adaptive

[21] See, also, Friedman (1977).

[22] While prices set in terms of dollars economize on the bookkeeping required to record relative prices, they only serve that purpose adequately in a monetary environment in which the evolution of the price level is predictable. There is then no "illusion" (confusion) about the relative price corresponding to a dollar price.

[23] Economists continue to divide over the issue of whether the central bank can exploit a Phillips curve relationship in order to mitigate large fluctuations in unemployment due to aggregate-demand shocks by increasing fluctuations in inflation. The converse case is that of mitigating large fluctuations in inflation due to inflation shocks by increasing fluctuations in an output gap. Goodfriend and King (1997) exposit the New Keynesian model in the monetarist spirit. The New Keynesian model as exposited by Clarida, Gali, and Gertler (1999) incorporates the assumption that the central bank can exploit a Phillips curve tradeoff in order to mitigate the effects on output of a real shock such as a markup or aggregate demand shock provided it follows a rule that commits it to returning inflation to a long-run target. The Clarida, Gali, and Gertler (1999) argument, however, does not address the issue of whether the central bank possesses the requisite knowledge of the structure of the economy (Friedman [1951] 1953; 1960). See the Appendix for skeptical comments on how well economists can estimate the structural coefficients of the New Keynesian Phillips curve.

expectations). The lag with which expectations adjusted to higher inflation could then explain the correlation between high (rising) inflation and low unemployment.

Friedman's formulation of the expectations-augmented Phillips curve, however, raised the theoretical possibility of long-run monetary nonneutrality. It appeared that the central bank could maintain the lower level of unemployment with ever-rising rates of inflation (the accelerationist hypothesis). For monetarists, the problem with that implication was that money was not necessarily neutral even in the long run in its influence on real variables (provided of course the central bank was willing to tolerate ever higher rates of inflation). As with the original Phillips curve, there appeared to be no unique equilibrium level of unemployment.

An answer to that problem led Robert Lucas to incorporate John Muth's idea of rational expectations into macroeconomics. Lucas ([1972] 1981) used the island paradigm employed by search models as a metaphor for incomplete information. He also imposed "rational expectations" in which the expectations of individuals are formed consistently with the structure of the economy and with the monetary policy followed by the central bank. Individuals on an island would alter output over confusion between a change in the overall island-wide price level and the relative price of their product. Within this model, Lucas stated the monetary neutrality proposition in a way that avoided the paradox of a central bank able to affect real output through systematic variation in the rate of inflation. The central bank could not permanently lower the unemployment rate through an ever-increasing inflation rate because the public would come to anticipate its actions and set prices in order to offset them. Such models incorporated what economists called the natural-rate/rational-expectations hypothesis.

Friedman had offered an explanation for the inverse correlations of the Phillips curve that predicted the disappearance of those correlations in response to sustained inflation. The stagflation of the United States in the 1970s supported that prediction. In reference to the Samuelson-Solow Phillips curve, Lucas and Sargent ([1978] 1981, 303) talked about "econometric failure on a grand scale." Lucas ([1973] 1981) argued that even the short-run tradeoff would tend to disappear as the variability of inflation increased.

Modigliani and Papademos (1975) offered the counterattack to the Friedman-Lucas critique. They pointed out that one could eliminate the empirically observed shifts in the Phillips curve by using first-differences of inflation. They then related first-differences in inflation to the difference in the unemployment rate and a benchmark value they termed the NIRU for "noninflationary rate of unemployment."

The NIRU (later called NAIRU for nonaccelerating inflation rate of unemployment) is the value of the unemployment rate for which inflation remains at its past value.[24] In practice, the estimated NAIRU is close to a slowly moving average of the past value of the unemployment rate.[25]

NAIRU models of inflation allowed for a long-run vertical Phillips curve. Apart from this assumption, however, they are in the tradition of the Samuelson-Solow Phillips curve. Originally, Keynesians adopted the Phillips curve because it supplied a connection between their IS-LM models, which were specified entirely for real variables, and inflation. The Phillips curve was an empirical relationship, not a theoretical one. It specified a relationship going from a real variable, unemployment, to a nominal variable, the rate of change of nominal wages (prices).[26] In NAIRU regressions, the unemployment rate relative to the NAIRU is the independent variable and inflation is the dependent variable. The central bank still possesses the ability to alter the rate of inflation through systematic control of a real variable, unemployment.

Keynesian economists argued that a Phillips curve with inflation in first differences represented a structural relationship that the central bank could use to smooth fluctuations in output around potential by imparting inverse fluctuations to changes in inflation.[27] The converse proposition came to be known as "flexible inflation targeting." That is, the central bank can eliminate an overshoot of inflation from target,

---

[24] Modigliani and Papademos suggested the archetypal NAIRU regression with inflation as the dependent variable and the unemployment rate and lagged inflation rates as independent variables. Estimation by constraining the coefficients on the lagged inflation terms to equal one allows calculation of the NAIRU. When inflation remains constant, the expectation of lagged inflation, given by the distributed lag of the inflation terms, equals the actual inflation rate. Consequently, the left-hand side variable (inflation) equals the right-hand side variable, expected inflation. The NAIRU then is the (negative) value of the constant term. That is, one solves the regression equation for the unemployment rate at which inflation equals expected inflation. Sargent ([1971] 1981) initiated a critique of this way of measuring expected inflation. In NAIRU regressions, the coefficients on the right-hand side of lagged inflation terms do not vary with changes in monetary policy. As a result, there is an inherent inertia in the expectations formation of the public that allows the policymaker to exploit a short-run Phillips curve tradeoff.

[25] King, Stock, and Watson (1995, 10) have found that "estimates of the NAIRU were very imprecise." Consistent with the monetarist hypothesis that monetary instability produces the inverse correlations of the Phillips curve, Dotsey, Fujita, and Stark (2011) found that the negative slope of the Phillips curve comes from recessions.

[26] The rationale for treating empirically estimated Phillips curves as structural derives from a generalization to the behavior of the price level of the way in which positive excess demand in individual markets produces relative price increases.

[27] King and Watson (1994) found a relationship between inflation and unemployment at business cycle frequencies, although not over lower frequency (trend) horizons. Their finding that inflation does not Granger cause (predict) unemployment, however, is not supportive of the idea that the central bank can manipulate inflation to control unemployment.

say, from an inflation shock, by raising the unemployment rate above its NAIRU value in a controlled way. The cost in terms of excess unemployment is given by the sacrifice ratio: the number of man-years of unemployment in excess of NAIRU the central bank must engineer to lower the inflation rate 1 percentage point.[28]

7. THE FIRST HALF OF THE SAMUELSON-SOLOW VAST EXPERIMENT

As noted above, the Phillips curve shifted upward in the 1970s. For example, in the 1950s, the unemployment rate among men 25 years and older averaged 3.5 percent. In the 1970s, it averaged 3.6 percent. In the 1950s, inflation (average, annualized monthly growth rates of CPI inflation) averaged 2.3 percent. In the 1970s, however, that figure rose to 7.5 percent. Similarly, annualized CPI inflation averaged over the first six months of 1964 was 0.85 percent while unemployment averaged 5.3 percent over this period. That figure was just slightly less than the 5.5 percent figure Samuelson and Solow had estimated as consistent with price stability. In contrast, for the 12-month period ending July 1971 (preceding the introduction of wage and price controls in August 1971), annualized monthly CPI inflation averaged 4.4 percent, while the unemployment rate averaged 5.8 percent.

In each case, the higher rate of inflation did not lower unemployment. Keynesians, however, attributed these upward shifts in inflation and the Phillips curve to cost-push shocks. In contrast, monetarists attributed them to shifts in expected inflation that frustrated the attempt to lower unemployment through aggregate-demand policies.

In 1970, 6 percent inflation accompanied 6 percent unemployment. Consistent with the prevailing Keynesian consensus, all but a minority of economists, mainly restricted to Chicago, Minneapolis, and the St. Louis Fed, interpreted the advent of this stagflation as a reflection of cost-push pressures that raised the level of the Phillips curve. In 1971, the Nixon administration turned to wage and price controls to restrain this presumed cost-push inflation and thus make way for an

[28] For example, David Stockton (Board of Governors of the Federal Reserve System 1989, 12) told the FOMC: "The sacrifice ratio is arrived at by dividing the amount of disinflation during a particular time period—measured in percentage points—into the cost of that disinflation—measured as the cumulative difference over the period between the actual unemployment rate and the natural rate of unemployment. Thus, it is a measure of the amount of excess unemployment over a year's time associated with each one percentage point decline in the inflation rate."

The staff reported that during the three post-Korean War disinflations, the sacrifice ratio was at or somewhat above 2. The exception was the period of price controls imposed in 1971.

expansionary monetary policy. Although those controls ended in 1974, the Carter administration resorted to various forms of incomes policies (see Hetzel [2008a, Chs. 8, 10, and 11]). These active attempts to control real output growth and unemployment while using incomes policies to control cost-push inflation created the experiment that Samuelson and Solow had talked about. The results contradicted the Keynesian assumption that policymakers could use aggregate-demand management in order to control real variables like unemployment in a systematic way and with a predictable cost in terms of inflation.

In the 1970s, Keynesian economists could see that supply shocks and a wage-price spiral drove inflation. The implication of rational expectations that a credible rule for monetary policy would shape the inflationary expectations of the public conformably with that rule appeared like an abstraction devoid of real-world relevance. It followed that a monetary policy objective of price stability that failed to accommodate inflation from nonmonetary causes would produce high unemployment. The following quotation from Paul Samuelson ([1979] 1986, 972) is representative of the times (see, also, Hetzel [2008a, Ch. 22]):

> Today's inflation is chronic. Its roots are deep in the very nature of the welfare state. [Establishment of price stability through monetary policy would require] abolishing the humane society [and would] reimpose inequality and suffering not tolerated under democracy. A fascist political state would be required to impose such a regime and preserve it. Short of a military junta that imprisons trade union activists and terrorizes intellectuals, this solution to inflation is unrealistic—and, to most of us, undesirable.

Samuelson's statement reflected the 1960s and 1970s Keynesian consensus that the behavior of the price level was determined by nonmonetary forces either having to do with real aggregate demand (demand pull) or with characteristics related to the lack of competitive markets such as the market power of large corporations and unions (cost push) (see, for example, Samuelson [1967]). The activist policy of aggregate-demand management combined with incomes policies of various degrees reflected this belief.[29]

On the international stage, Keynesian policy prescriptions played out in countries that pegged their exchange rates to the dollar as part of the Bretton Woods system. As reflected in the Keynesian spirit of the time, countries with pegged exchange rates also followed policies of aggregate-demand management intended to maintain full employment

[29] The term "incomes policies" refers to any government intervention into the wage and price setting of the private sector. Wage and price controls are an extreme version.

(see Capie [2010] for the United Kingdom case). As Friedman ([1953b] 1953) had predicted, these countries had to resort to capital controls as well as wage and price controls in order to reconcile an exchange rate peg with an unwillingness to allow their internal price levels to adjust in order to vary the real terms of trade to achieve balance of payments equilibrium. In 1973, the Bretton Woods system of pegged exchange rates collapsed (Hetzel 2008a, Ch. 9).

By the end of the 1970s, the experiment with activist monetary policy concluded with double-digit inflation accompanied by cyclical instability. However, as noted above, despite the unusual clarity about policy, extraneous forces always prevent these episodes from offering the kind of certitude as a controlled experiment in the physical sciences. The issue remains whether activist monetary policy produced this result or whether a series of adverse inflation shocks overwhelmed the stabilizing properties of activist policy.[30] Velde (2004) characterized the issue as one of bad hand (inflation shocks) or bad play (destabilizing monetary policy). In early 1979, the United States could have continued the experiment with activist monetary policy reinforced by a return to wage and price controls. However, a change in the political landscape with the election of Ronald Reagan as president, combined with the way in which individuals occasionally change the course of events in the form of Paul Volcker as FOMC chairman, gave the United States a very different kind of monetary experiment.[31]

8. THE SECOND PART OF THE VAST EXPERIMENT

The back-to-back experience of the Great Depression with World War II created the Keynesian consensus. The back-to-back experience in the 1970s of an activist policy directed toward maintaining low, stable unemployment and the policy in the 1980s and 1990s of restoring price stability through restoring nominal expectational stability flipped the professional consensus. The profession came to see inflation as a monetary phenomenon. Also, countries realized that if they were to control their own price levels, they had to abandon fixed exchange rates in favor of floating exchange rates in order to gain control over money

[30] Gordon (1985) and Sims and Zha (2006) emphasized the importance of inflation shocks. Sims and Zha (2006, 54) argued that "the differences among [monetary policy] regimes are not large enough to account for the rise, then decline, in inflation of the 1970s and 1980s." Blinder (1987, 133) wrote: "The fact is that, the Lucas critique notwithstanding, the Phillips curve, once modified to allow for supply shocks ... has been one of the best-behaved empirical regularities in macroeconomics...."

[31] On the political economy of the late 1970s, see Hetzel (2008a, Ch. 12).

creation. Having floated their exchange rates, countries realized that they had to leave the control of inflation to the central bank.

The second part of the "vast experiment" was then the effort by the Volcker and Greenspan FOMCs to restore the nominal expectational stability lost in the preceding stop-go era (Hetzel 2008b). The Volcker-Greenspan FOMCs discarded the idea of measuring the level of idle resources (the output gap). Instead, they moved the funds rate in a persistent way designed to counter sustained changes in the rate of resource utilization. That is, they removed the measurement error inherent in trying to measure the level of idle resources by focusing on changes in the degree of resource utilization (Orphanides and Williams 2002). Given the desire to restore credibility in instances of sustained increases in the rate of resource utilization, the Fed watched bond markets for evidence that the "bond market vigilantes" were satisfied that increases in the funds rate would cumulate to a sufficient degree in order to prevent a revival of inflation. In response to inflation scares, the FOMC raised the funds rate more aggressively (Goodfriend 1993).

The willingness of the FOMC to move the funds rate in a sustained way made it clear to markets that it had abandoned the prior practice of inferring the thrust of monetary policy from a "high" or "low" level of short-term interest rates. That is, the FOMC did not back off from changes in the funds rate when the funds rate reached a "high" or "low" level. These procedures, termed "lean-against-the-wind with credibility" by Hetzel (2008a), removed the cyclical inertia from interest rates (see Hetzel [2008a, Chs. 14, 15, 21, and 22]). Equivalently, the discipline they imposed in removing cyclical inertia from funds rate changes prevented attempts to use Phillips curve tradeoffs to achieve macroeconomic objectives.

The demonstration that the Fed could maintain low, stable inflation without incurring the cost of recurrent bouts of high unemployment weakened the Keynesian consensus. The economics profession became receptive to replacement of the IS-LM model with what would become, in time, the New Keynesian model. In the Great Inflation, Keynesians had fleshed out the IS-LM model with explanations of inflation that turned on a wage-price spiral propelled by expectations of inflation untethered by monetary policy. They also assumed the existence of negative output gaps persisting over many years arising from the weak equilibrating properties of the price system. The New Keynesian model challenged the self-evident descriptive realism of such assumptions with incorporation of rational expectations and an inner real-business-cycle core in which the price system worked well to maintain macroeconomic equilibrium.

The traditional Keynesian Phillips curve with inflation generated by the momentum of lagged inflation and an output gap measured as cyclical deviations of output from a smooth trend ceded place to the New Keynesian Phillips curve. The forward-looking agents posited by the New Keynesian model base their behavior not only on the current policy actions of the central bank but also on the way in which the central bank's systematic behavior shapes the policy actions it takes in the future in response to incoming data on the economy. As a result, contemporaneous inflation (current price-setting behavior) depends on the expectation of future inflation, which depends on the rule the central bank implements.

9. THE GREAT DEBATE WILL CONTINUE

The recent Great Recession has weakened the New Keynesian consensus described above, at least in the Goodfriend-King (1997) version in which the optimal policy for the central bank is to stabilize the price level and thereby allow the real-business-cycle core of the economy to control the behavior of the real economy. To a significant extent, both popular and much professional commentary have reverted to the historical "default option" for explanations of the business cycle—the "imbalances" model (Hetzel 2012, Ch. 2). The business cycle is self-generating because imbalances accumulate during periods of expansion. At some point, the extent of maladjustments cumulates to the point at which a correction becomes inevitable. The economy must then endure a period of purging of the economic body.

In financial markets, these imbalances appear as credit cycles. In periods of economic expansion, investors become overly optimistic about the future. They take on debt and push asset prices to levels not supported by the underlying productive capacity of the assets. Inevitably, these asset bubbles burst. Investors find themselves with too much debt. A long, painful process of deleveraging ensues in which economic activity is depressed. When this process works its way out, recovery can begin. Once again, the process of swings in investor sentiment from unfounded optimism to unfounded pessimism begins. Commentary in this vein on the Great Recession has focused on an asset bubble in the housing market made possible by expansionary monetary policy in the years preceding 2008.

In order to move beyond the "descriptive reality" of these age-old explanations of the business cycle based on the correlation that in economic booms asset prices rise and debt increases while in recessions asset prices decline and debt declines, one needs a model and plausible exogenous shocks. The Keynesian model with its swings in animal

spirits among investors that overwhelm the stabilizing properties of the price system was an attempt to construct such a model. In the spirit of this article, how will economists test the imbalances hypothesis or Keynesian versions of it against the monetarist hypothesis that highlights as the precipitating factor in recessions central bank interference with the operation of the price system?

To recapitulate the discussion of methodology of Section 1, there will be a multitude of models assuming different shocks and different structures of the economy and frictions that can explain historical time series and, a fortiori, particular events like the Great Recession. It is thus improbable that economists will ever reach consensus over the cause of a particular recession. However, scholarly debate will return to the pattern of asking how well a particular recession like the Great Recession fits into one of the alternative frameworks that explain the recurrent phenomenon of cyclical fluctuations. Economists will continue running horse races among models based on the entire historical record. Using models based on microeconomic foundations, they will ask whether the implications of the model adequately explain correlations in the entire historical record that are robust in that the correlations persist over time and across countries, that is, in a variety of circumstances. The latter characteristic is the social sciences version of the controlled experiment in the physical sciences.

Consider the correlation between monetary and real instability. The monetarist hypothesis is that, to a significant degree, causation runs from monetary to real instability. In the world of Milton Friedman, prior to 1981, given the existence of a monetary aggregate (M1), which was interest insensitive and stably related to nominal output (GDP), the robust correlation was that monetary decelerations preceded business cycle peaks. Furthermore, the central bank behavior that accompanied those monetary decelerations plausibly produced changes in nominal money originating independently of changes in real money demand. The robustness of this generalization across countries and across time reduces the possibility that it reflects causation produced by some third variable so that real instability arises independently of monetary instability. Of course, no controlled experiment produced these correlations. The hypothesis that monetary instability produces real instability has to be put into a form in which it yields testable predictions about the future.

Because of the disappearance since 1981 of a monetary aggregate like M1 that is useful as a predictor of nominal GDP, it is necessary to refocus the search for robust correlations based on the monetarist hypothesis that monetary disorder originates in central bank interference with the operation of the price system. Reformulated in this spirit,

the monetarist hypothesis receives support from the continuance of the central bank behavior associated with the monetary decelerations preceding business cycle peaks in the pre-1981 period.

What is this central bank behavior? In the post-World War II period, when the Fed became concerned about inflation, it first raised interest rates and then, out of a concern not to exacerbate inflationary expectations, introduced inertia into the downward adjustment of interest rates when the economy weakened (Hetzel 2012, Ch. 8).[32] Although the Fed did not employ the language of tradeoffs, these attempts to exploit a Phillips curve relationship by allowing a negative output gap to develop have constituted a reliable leading indicator of recession (Romer and Romer 1989; Hetzel 2008a, Chs. 23–25; Hetzel 2012, Chs. 6–8). The same empirical regularity existed in the pre-World War II period, but the Fed raised rates and then introduced inertia into the downward adjustment of interest rates while the economy weakened not out of concern for inflation but out of concern that the level of asset prices reflected a speculative asset bubble.

Hetzel (2009, 2012, 2013b) argues that the Great Recession fits into this monetarist characterization of central bank behavior associated with recessions. The persistent inflation shock that began in summer 2004 intensified in summer 2008 and pushed headline inflation well above core inflation and central bank inflation targets. That inflation shock created a moderate recession by dampening growth of real disposable income. Moderate recession turned into severe recession in summer 2008 when central banks either raised interest rates (the European Central Bank) or left them unchanged as economic activity weakened (the Fed). The attempt to create a negative output gap to bend inflation down mirrored the stop phases of the earlier stop-go monetary policy.

10. TESTING THEORIES OF THE BUSINESS CYCLE

In the absence of consensus within the economics profession over the causes of the business cycle, popular commentary fills the void with explanations based on descriptive reality. That verbiage is inevitable given the importance of phenomena like cyclical fluctuations in unemployment. However, economists do possess a methodology for learning and will make progress in understanding the causes of the business

[32] The exceptions are especially important for evaluating robust correlations. Prior to the April 1960 business cycle peak, the FOMC raised rates and then maintained them despite a weakening economy out of a concern not for inflation but rather out of concern for a deficit in international payments and gold outflows (Hetzel 1996; 2008a, 52–5).

cycle. In this respect, the stumbling, painful, and ongoing process of the central bank learning how to manage the fiat money regime that replaced the earlier commodity standards remains a still under-investigated source of the semi-controlled experiments required to extract causation from correlation.

APPENDIX: RECENT WORK ON THE PHILLIPS CURVE

Little in the work on the New Keynesian Phillips curve (NKPC) challenges the Friedman assertion that policymakers lack sufficient information about the structure of the economy in order to implement an activist monetary policy. As summarized by Hornstein (2008), the results of empirical estimation of the NKPC offer little useful information for the policymaker interested in exploiting a Phillips curve tradeoff. For example, Hornstein (2008, 305) comments:

> Nason and Smith [2008] also discuss the finding that the estimated coefficient on marginal cost tends to be small and barely significant. This is bad news for the NKPC as a model of inflation and for monetary policy.

The coefficient on real marginal cost referred to summarizes the real-nominal interaction implied by the nominal price stickiness in the New Keynesian model. As implied in the above quotation, econometric estimation provides no practical guidance for monetary policy procedures based on Phillips curve tradeoffs.

Hornstein elucidates the reasons for this lack of guidance in his discussion of Schorfeide (2008). Estimation of the NKPC through single-equation methods founders on the seemingly technical but fundamental issue of the lack of plausible instruments useful for forecasting inflation, while at the same time being unrelated to the other variables in the Phillips curve and macroeconomic shocks. Everything in macroeconomics is endogenously determined. The alternative is to treat the elements in the NKPC, like real marginal cost, as "latent variables," that is, variables not observable but constructed from the equations of a complete model. The problem then is that different models yield different measures and there is no consensus on the true model (the model useful for the analysis of policy).

Given a model with a NKPC, Schmitt-Grohé and Uribe (2008) conduct a normative exercise evaluating different monetary policy rules.

However, as Hornstein (2008, 307) notes, with "no agreement on how substantial nominal rigidities are" it is hard to know how useful such exercises are for policy. For example, the authors make use of a Taylor rule, which assumes that the central bank can respond directly to misses in its inflation target without destabilizing the economy. In actual practice, the assumption is that in response to such a miss, the central bank can create a controlled negative output gap (increase firms' markups in order to eliminate the miss). The whole issue then reemerges of whether central banks can control inflation through exploiting a Phillips curve tradeoff. The Lucas-Friedman contention that attempts by the central bank to exploit real-nominal relationships destabilize the economy remains a live issue.

The econometric difficulties highlighted by Hornstein (2008) turn ultimately on the issue of identification, both of shocks and of structural relationships. That fact suggests that in future research the profession should revive the monetarist identification scheme implicit in the work of King (2008), who uses historical narrative to isolate the monetary policy experiments conducted by the regime changes of central banks (see, also, Hetzel [2008a, 2012]).

REFERENCES

Ando, Albert, and Franco Modigliani. 1965. "The Relative Stability of Monetary Velocity and the Investment Multiplier." *American Economic Review* 55 (September): 693–728.

Barnett, William A. 1982. "Divisia Indices." In *Encyclopedia of Statistical Sciences, Vol. 2*, edited by Samuel Kotz and Norman L. Johnson. New York: Wiley.

Blinder, Alan S. 1987. "Keynes, Lucas, and Scientific Progress." *American Economic Review* 77 (May): 130–6.

Board of Governors of the Federal Reserve System. 1989. Transcripts of the Federal Open Market Committee, December 18.

Breit, William, and Roger L. Ransom. 1982. *The Academic Scribblers*. Chicago: The Dryden Press.

Capie, Forrest. 2010. *The Bank of England: 1950s to 1979*. Cambridge: Cambridge University Press.

Chari, V. V., Patrick J. Kehoe, and Ellen R. McGrattan. 2009. "New Keynesian Models: Not Yet Useful for Policy Analysis." *American Economic Journal: Macroeconomics* 1 (January): 242–66.

Clarida, Richard, Jordi Gali, and Mark Gertler. 1999. "The Science of Monetary Policy: A New Keynesian Perspective." *Journal of Economic Literature* 37 (December): 1,661–707.

Dotsey, Michael, Shigeru Fujita, and Tom Stark. 2011. "Do Phillips Curves Conditionally Help to Forecast Inflation?" Federal Reserve Bank of Philadelphia Working Paper 11-40 (September).

Friedman, Milton. [1951] 1953. "The Effects of a Full-Employment Policy on Economic Stability: A Formal Analysis." In *Essays in Positive Economics*, edited by Milton Friedman. Chicago: The University of Chicago Press, 117–32.

Friedman, Milton. [1953a] 1953. "The Methodology of Positive Economics." In *Essays in Positive Economics*, edited by Milton Friedman. Chicago: The University of Chicago Press, 3–46.

Friedman, Milton. [1953b] 1953. "The Case for Flexible Exchange Rates." In *Essays in Positive Economics*, edited by Milton Friedman. Chicago: The University of Chicago Press, 157–203.

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

Friedman, Milton. [1968] 1969. "The Role of Monetary Policy." In *The Optimum Quantity of Money and Other Essays*, edited by Milton Friedman. Chicago: Aldine, 95–110.

Friedman, Milton. 1977. "Nobel Lecture: Inflation and Unemployment." *Journal of Political Economy* 85 (June): 451–72.

Friedman, Milton, and Anna J. Schwartz. 1963a. "Money and Business Cycles." *Review of Economics and Statistics* 45 (February): 32–64.

Friedman, Milton, and Anna J. Schwartz. 1963b. *A Monetary History of the United States, 1867–1960.* Princeton, N.J.: Princeton University Press.

Friedman, Milton, and Anna J. Schwartz. 1991. "Alternative Approaches to Analyzing Economic Data." *American Economic Review* 81 (March): 39–49.

Friedman, Milton, and David Meiselman. 1963. "The Relative Stability of Monetary Velocity and the Investment Multiplier in the United States, 1897–1958." In *Stabilization Policies.* Englewood Cliffs, N.J.: Prentice-Hall, 165–268.

Goodfriend, Marvin. 1993. "Interest Rate Policy and the Inflation Scare Problem." Federal Reserve Bank of Richmond *Economic Quarterly* 79 (Winter): 1–24.

Goodfriend, Marvin, and Robert G. King. 1997. "The New Neoclassical Synthesis and the Role of Monetary Policy." In *NBER Macroeconomics Annual*, edited by Ben S. Bernanke and Julio Rotemberg. Cambridge, Mass.: MIT Press, 231–96

Gordon, Robert J. 1985. "Understanding Inflation in the 1980s." *Brookings Papers on Economic Activity* 16 (1): 263–302.

Hayek, Friedrich A. 1945. "The Use of Knowledge in Society." *American Economic Review* 35 (September): 519–30.

Hetzel, Robert L. 1985. "The Rules versus Discretion Debate over Monetary Policy in the 1920s." Federal Reserve Bank of Richmond *Economic Review* 71 (November/December): 3–14.

Hetzel, Robert L. 1996. "Sterilized Foreign Exchange Intervention: The Fed Debate in the 1960s." Federal Reserve Bank of Richmond *Economic Quarterly* 82 (Spring): 21–46.

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

Hetzel, Robert L. 2008a. *The Monetary Policy of the Federal Reserve: A History*. New York: Cambridge University Press.

Hetzel, Robert L. 2008b. "What Is the Monetary Standard, Or, How Did the Volcker-Greenspan FOMC's Tame Inflation?" Federal Reserve Bank of Richmond *Economic Quarterly* 94 (Spring): 147–71.

Hetzel, Robert L. 2009. "Monetary Policy in the 2008–2009 Recession." Federal Reserve Bank of Richmond *Economic Quarterly* 95 (Spring): 201–33.

Hetzel, Robert L. 2012. *The Great Recession: Market Failure or Policy Failure?* New York: Cambridge University Press.

Hetzel, Robert L. 2013a. "The Great Inflation of the 1970s." In *The Handbook of Major Events in Economic History*, edited by Randall Parker and Robert Whaples. New York: Routledge, 223–38.

Hetzel, Robert L. 2013b. "ECB Monetary Policy in the Recession: A New Keynesian (Old Monetarist) Critique." Federal Reserve Bank of Richmond Working Paper 13-07R (July).

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

Hornstein, Andreas. 2008. "Introduction to the New Keynesian Phillips Curve." Federal Reserve Bank of Richmond *Economic Quarterly* 94 (Fall): 301–9.

Humphrey, Thomas M. 1986. "From Trade-offs to Policy Ineffectiveness: A History of the Phillips Curve." Federal Reserve Bank of Richmond *Monograph* 1986 (October).

Keynes, John Maynard. [1936] 1973. *The Collected Writings of John Maynard Keynes, vol. 7*. London: The Macmillan Press.

King, Robert G. 2008. "The Phillips Curve and U.S. Macroeconomic Policy: Snapshots, 1958–1996." Federal Reserve Bank of Richmond *Economic Quarterly* 94 (Fall): 311–59.

King, Robert G., and Mark W. Watson. 1994. "The Post-war U.S. Phillips Curve: A Revisionist Econometric History." Carnegie-Rochester Conference Series on Public Policy 41 (December): 157–219.

King, Robert G., James H. Stock, and Mark W. Watson. 1995. "Temporal Instability of the Unemployment-Inflation Relationship." Federal Reserve Bank of Chicago *Economic Perspectives* 19 (May/June): 2–12.

Koopmans, Tjalling. 1947. "Measurement without Theory." *The Review of Economic Statistics* 28 (August): 161–72.

Lucas, Robert E., Jr. [1972] 1981. "Expectations and the Neutrality of Money." In *Studies in Business-Cycle Theory*, edited by Robert E. Lucas, Jr. Cambridge, Mass.: The MIT Press, 66–89.

Lucas, Robert E., Jr. [1973] 1981. "Some International Evidence on Output-Inflation Tradeoffs." In *Studies in Business-Cycle Theory*, edited by Robert E. Lucas, Jr. Cambridge, Mass.: The MIT Press, 131–45.

Lucas, Robert E., Jr., and Thomas J. Sargent. [1978] 1981. "After Keynesian Macroeconomics." In *Rational Expectations and Econometric Practice, vol 1*, edited by Robert E. Lucas, Jr., and Thomas J. Sargent. Minneapolis: The University of Minnesota Press, 295–319.

Modigliani, Franco, and Lucas Papademos. 1975. "Targets for Monetary Policy in the Coming Year." *Brookings Papers on Economic Activity* 6 (1): 141–63.

Nason, James M., and Gregor W. Smith. 2008. "The New Keynesian Phillips Curve: Lessons from Single-Equation Econometric Estimation." Federal Reserve Bank of Richmond *Economic Quarterly* 94 (Fall): 361–96.

Orphanides, Athanasios, and John C. Williams. 2002. "Robust Monetary Policy Rules with Unknown Natural Rates." *Brookings Papers on Economic Activity* 33 (2): 63–145.

Phillips, A. W. 1958. "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957." *Economica* 25 (November): 283–99.

Romer, Christina D., and David H. Romer. 1989. "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz." In *NBER Macroeconomics Annual 1989, vol. 4*, edited by Olivier Jean Blanchard and Stanley Fischer. Cambridge, Mass.: The MIT Press, 121–84.

Samuelson, Paul A. 1967. *Economics: An Introductory Analysis*. New York: McGraw-Hill Book Company.

Samuelson, Paul A. [1979] 1986. "Living with Stagflation." In *The Collected Scientific Papers of Paul A. Samuelson, vol. 5, no. 379*, edited by Kate Crowley. Cambridge, Mass.: The MIT Press, 972.

Samuelson, Paul, and Robert Solow. [1960] 1966. "Analytical Aspects of Anti-Inflation Policy." In *The Collected Scientific Papers of Paul A. Samuelson, vol. 2, no. 102*, edited by Joseph Stiglitz. Cambridge, Mass.: The MIT Press, 1,336–53.

Sargent, Thomas J. [1971] 1981. "A Note on the 'Accelerationist' Controversy." In *Rational Expectations and Econometric Practice, vol 1*, edited by Robert E. Lucas, Jr., and Thomas J. Sargent. Minneapolis: The University of Minnesota Press, 33–8.

Schmitt-Grohé, Stephanie, and Martín Uribe. 2008. "Policy Implications of the New Keynesian Phillips Curve." Federal Reserve Bank of Richmond *Economic Quarterly* 94 (Fall): 435–65.

Schorfheide, Frank. 2008. "DSGE Model-Based Estimation of the New Keynesian Phillips Curve." Federal Reserve Bank of Richmond *Economic Quarterly* 94 (Fall): 397–434.

Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica* 48 (January): 1–48.

Sims, Christopher A., and Tao Zha. 2006. "Were There Regime Switches in U.S. Monetary Policy." *American Economic Review* 96 (March): 54–81.

Stockman, Alan C. 1996. *Introduction to Economics*. Orlando: The Dryden Press.

Velde, François. 2004. "Poor Hand or Poor Play? The Rise and Fall of Inflation in the U.S." Federal Reserve Bank of Chicago *Economic Perspectives* 28 (Q1): 34–51.

Economic Quarterly—Volume 99, Number 2—Second Quarter 2013—Pages 117–141

Federal Reserve Interdistrict Settlement

Alexander L. Wolman

The Interdistrict Settlement Account (ISA) is used to keep track of movements in assets and liabilities across Federal Reserve Banks within the Federal Reserve System. To the extent that the independent financial status of individual Federal Reserve Banks is meaningful, the ISA is the means by which each Bank grants credit to the other Banks in the System. Even if one views financial independence as more apparent than real, the behavior of individual Reserve Bank balance sheet components, including ISA, can shed light on ongoing financial developments in the economy. This article provides an introduction to the ISA and traces the behavior of ISA and some other components of Reserve Bank balance sheets during the Great Recession and the financial crisis. In addition, it provides some speculative discussion of how Reserve Bank balance sheets could be informative about economic conditions as the Fed exits from unconventional monetary policy.

The ISA may seem like an obscure topic. However, in 2012 the European debt crisis led to much discussion of the TARGET2 system, which is—loosely—Europe's analogue to the combination of ISA and the Fedwire funds transfer system (see Cecchetti, McCauley, and McGuire [2012], Whelan [2012], and the references therein). Discussions about TARGET2 often included comparisons—some of them shaky—to ISA, drawing attention to the fact that there were few sources available describing ISA to the lay public.[1] In attempting to help fill

The author is grateful to Ceci Adams for her patient explanations of ISA accounting, and to Huberto Ennis, Peter Garber, Bob Hetzel, J.P. Koning, Marisa Reed, Karl Rhodes, and John Weinberg for comments and discussions. The views in this article are the author's. They do not represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: alexander.wolman@rich.frb.org.

[1] Lubik and Rhodes (2012) provide a concise summary of ISA in their essay on TARGET2. Koning (2012) provides a more detailed discussion of ISA, including a

that void, this article also discusses two important ways in which ISA differs from TARGET2.

Monetary policy in the United States is implemented primarily by the Federal Reserve Bank of New York. For example, the securities purchases that comprise the Federal Open Market Committee's (FOMC) large-scale asset purchase programs (LSAPs) are conducted by the New York Fed. However, securities purchased by the New York Fed are apportioned the same day to all 12 Federal Reserve Banks, and there is an annual rebalancing of Federal Reserve Bank balance sheets. Both the apportionment and rebalancing involve use of the ISA, and in recent years these have been main drivers of the ISA. As such we provide a relatively detailed discussion of both topics. Apportionment assures that all 12 Federal Reserve Banks are effectively equal stakeholders in monetary policy operations; the New York Fed simply acts as agent for the other 11 Banks. Rebalancing, in turn, assures that over time the securities are held by Reserve Banks in rough proportion to the liabilities that have been issued by those Reserve Banks.

There is one authoritative source for the ISA, the Federal Reserve's Financial Accounting Manual (FAM). While the FAM is publicly available, it is written for users and not for the interested public. This article is not a substitute for the FAM, but should provide an accessible introduction to the ISA for readers without the time or inclination to delve into the FAM. In that context, it is important to stress that the language and terminology used here conflict at times with the language used in the FAM. Note in particular that ISA balances will be referred to throughout as an asset that can enter with a positive or negative sign on Federal Reserve Bank balance sheets; this is the same convention used in the Federal Reserve Board's H.4.1 release, which is the source for most of the data used in the article.

1. THE INTERDISTRICT SETTLEMENT ACCOUNT: OVERVIEW AND EXAMPLES

Each of the 12 Federal Reserve Banks has its own balance sheet. The assets on a Federal Reserve Bank's balance sheet currently consist primarily of securities allocated to the bank by the New York Fed. The liabilities consist mainly of Federal Reserve notes in circulation (paper currency) and reserve accounts of banks located in the Reserve District. Many transactions that affect a Reserve Bank's balance sheet involve only the Reserve Bank and a commercial bank. For example, if a

history of clearing and settlement across Federal Reserve Banks. Eichengreen, Mehl, and Chitu (2013) also discuss that history—see Section 2.

Table 1 T-Accounts for Commercial Banks, Check Clearing Example

Paying Commercial Bank ("Bank A")	
Assets	**Liabilities**
−$1 million, Reserve account (at Richmond Fed)	−$1 million, customer deposits
Receiving Commercial Bank ("Bank B")	
Assets	**Liabilities**
+$1 million, Reserve account (at Atlanta Fed)	+$1 million, customer deposits

commercial bank withdraws currency from the Federal Reserve Bank of Richmond, there is an increase in the Richmond Fed's net Federal Reserve notes outstanding, and an offsetting decrease in reserves (denoted "other deposits held by depository institutions" on the balance sheet as represented by the H.4.1 release); and if the Richmond Fed makes a discount window loan to a commercial bank (necessarily in its district), then there is an increase in the Richmond Fed's loan assets, and an increase in its reserve liabilities. Other transactions, however, affect the balance sheets of more than one Federal Reserve Bank. The ISA is a line item on the asset side of each Federal Reserve Bank's balance sheet that is used to account for transactions across Federal Reserve Banks. It can be negative or positive for a single Reserve Bank and always sums to zero across the 12 Reserve Banks.[2]

The ISA can be best understood through examples of different types of transactions. Transactions that are initiated by commercial banks are relatively easy to explain, whereas transactions that are undertaken by the Federal Reserve Bank of New York as part of the Fed's monetary or credit policy implementation are more complicated and will lead us into the discussion of allocation/apportionment in the next section. For each example, we will provide both a verbal discussion and a summary using T-accounts.

Consider first a stylized situation where customers of a commercial bank in the Fifth Federal Reserve District (Richmond) write checks to customers of a commercial bank in the Sixth Federal Reserve District (Atlanta) in the net amount of $1 million. The paying commercial bank will see its reserve account at the Richmond Fed (an asset on

[2] The current system for accommodating deficits and surpluses across Federal Reserve Districts dates back to 1975. See Eichengreen, Mehl, and Chitu (2013) for a description and analysis of the pre-1975 system, focusing on the period from 1913 to 1960.

Table 2 T-Accounts for Federal Reserve Banks, Check Clearing Example

Paying Federal Reserve Bank (Richmond)	
Assets	**Liabilities**
−$1 million, ISA balances	−$1 million, Bank A reserve account
Receiving Federal Reserve Bank (Atlanta)	
Assets	**Liabilities**
+$1 million, ISA balances	+$1 million, Bank B reserve account

the commercial bank's balance sheet) reduced by $1 million, and it will see its customers' deposits (a liability) reduced by $1 million. The receiving commercial bank will see corresponding increases in its reserve account at the Atlanta Fed and in its customers' deposits. Just as both commercial banks have balanced changes in assets and liabilities, so do both Federal Reserve Banks. The Richmond Fed's reserve account liabilities decrease by $1 million, the Atlanta Fed's reserve account liabilities increase by $1 million, and the offsetting changes on the asset side of the Reserve Banks' balance sheets occur through the ISA. Because the Richmond Fed is effectively making a payment to the Atlanta Fed, its ISA balance (an asset) falls by $1 million, and the Atlanta Fed's ISA balance rises by $1 million. Tables 1 and 2 show the relevant T-accounts. If ISA did not exist, there are two possibilities for how to account for this transaction (ignoring legal issues). One possibility is that securities or other assets could be transferred from the Richmond Fed to the Atlanta Fed.[3] Alternatively, the balance sheets of the Federal Reserve Banks could be consolidated, so that the transaction would simply involve a relabeling of accounts with the single Federal Reserve Bank. We will not go through these alternatives for the other examples below, but the reader should keep in mind that similar reasoning applies.

Next, consider delivery of $1 million of new currency (bills) to the Federal Reserve Bank of New York (for example), where the new currency is designated as issued by the Federal Reserve Bank of San Francisco.[4] In this case, the Federal Reserve Bank of San Francisco's

[3] Eichengreen, Mehl, and Chitu (2013) discuss how, before 1975, instead of ISA there was a combination of settlement through transfer of gold certificates (to be discussed below) and discretionary "mutual assistance" among Reserve Banks.

[4] All currency is designated as issued by one of the 12 Federal Reserve Banks, and is marked with the corresponding district number and letter. As this example suggests, however, currency does not necessarily enter circulation in the district through which it is officially issued.

Table 3 T-Accounts for Federal Reserve Banks, New Currency Example

Federal Reserve Bank of San Francisco	
Assets	**Liabilities**
+$1 million, ISA balances	+$1 million, Federal Reserve Notes outstanding

Federal Reserve Bank of New York	
Assets	**Liabilities**
−$1 million, ISA balances	−$1 million, Notes held by Federal Reserve Banks

liabilities increase by $1 million ("Federal Reserve notes outstanding" on the H.4.1 release) and the Federal Reserve Bank of New York's liabilities decrease by $1 million ("Notes held by Federal Reserve Banks" on the H.4.1).[5] Of course, both Banks must have an offsetting balance sheet change, and these involve the ISA: The San Francisco Fed's ISA balance increases by $1 million, and the New York Fed's ISA balance decreases by $1 million. The T-accounts are trivial in this case, shown in Table 3. In effect, the New York Fed is purchasing currency from the San Francisco Fed using its ISA account.

We move now to transactions related to the implementation of monetary or credit policy. These transactions are typically initiated by the Federal Reserve Bank of New York, and thus at first only impact the New York Fed's balance sheet.[6] However, according to the policies set forth in the FAM, the associated balance sheet changes are apportioned on a daily basis to all 12 Federal Reserve Banks.

Consider first a typical asset purchase that affects the domestic portfolio of the System Open Market Account (SOMA). The Fed's ongoing large-scale asset purchases fall into this category, so we will use a specific example of one of these purchases. On December 27, 2012, the Federal Reserve Bank of New York purchased $4.614 billion of Treasury securities from the primary dealers who serve as trading counterparties with the New York Fed.[7] These purchases settled on

[5] "Notes held by Federal Reserve Banks" appears on the liability side of each Federal Reserve Bank's balance sheet. However, on the liability side it is *deducted* from the value of Federal Reserve notes outstanding. Thus, if a Reserve Bank has $10 billion in notes outstanding, and holds $100 million of notes in its vaults, then its consolidated liability for these items is $9.9 billion.

[6] Some forms of credit policy, for example the Term Auction Facility, initially hit all the Reserve Bank balance sheets, to the extent that commercial banks in all 12 Districts borrow at the auction. In contrast, the Maiden Lane facilities involved only the New York Fed's balance sheet.

[7] A complete list of purchases is available at www.newyorkfed.org/markets/pomo/display/index.cfm.

Table 4 SOMA Portfolio Allocation Percentages

	Domestic		Foreign	
District	**2012**	**2011**	**2012**	**2011**
Boston	2.429	2.459	3.506	3.456
New York	56.065	46.504	32.258	28.963
Philadelphia	3.306	3.426	8.674	9.686
Cleveland	2.542	2.701	7.393	7.418
Richmond	7.117	11.549	20.685	20.505
Atlanta	6.029	7.434	5.718	5.731
Chicago	5.548	5.939	2.668	2.534
St. Louis	1.563	1.893	0.818	0.815
Minneapolis	0.909	1.537	0.408	3.089
Kansas City	2.009	2.660	0.995	0.900
Dallas	3.885	3.955	1.602	1.515
San Francisco	8.596	9.944	15.277	15.388
System Total	100	100	100	100

December 28, which means that on December 28 the Federal Reserve Bank of New York's securities holdings (an asset) increased by $4.614 billion. The primary dealers were paid for these securities by credits to their accounts in reserve-holding banks; thus, the New York Fed's reserve liabilities increased by $4.614 billion.[8] Subsequently, but still on December 28, the $4.614 billion increase in securities holdings was apportioned to all 12 Federal Reserve Banks according to the percentages listed in the second column of Table 4. How those percentages are determined will be discussed in detail in the next section; the procedure is complicated, but loosely it tends to assign higher percentages to Reserve Banks with a higher percentage of currency outstanding and deposit liabilities. The reduction in the New York Fed's securities holdings and the increases in the other Reserve Banks' securities holdings were offset by increases in New York's ISA balance and decreases in the other Banks' ISA balances. Again, it is as if the other 11 Federal Reserve Banks purchased securities from the New York Fed using their ISA accounts. Table 5 puts this example in T-account form, for the New York Fed and the Richmond Fed. New York has two steps; in the first step it receives all the securities, and in the second step it apportions 43.935 percent of the securities to the other 11 Banks. In the apportionment step, 7.117 percent of the securities are apportioned to Richmond.

[8] In principle, a primary dealer's deposit account could be with a bank located outside the New York Federal Reserve District. In this case ISA would be involved in the initial transaction. For simplicity we assume that the primary dealer's bank has a reserve account with the New York Fed.

Table 5 T-Accounts for Federal Reserve Banks, Asset Purchase Example

	Federal Reserve Bank of New York	
	Assets	**Liabilities**
Step 1	+$4.614 billion securities	+$4.614 billion commercial bank deposits (reserves)
Step 2	−$2.027 billion securities	—
Step 3	+$2.027 ISA balances	—
	Federal Reserve Bank of Richmond	
	Assets	**Liabilities**
	+$328 million securities	—
	−$328 million ISA balances	—

A similar process occurs for foreign-currency denominated assets in the SOMA portfolio, but the apportionment uses percentages based on member bank capital in each district. Apportionment will be discussed in more detail below. An example of a foreign-currency denominated transaction occurred the week of August 15, 2012, when the European Central Bank (ECB) drew on its swap line with the Federal Reserve Bank of New York by $7 billion; the swap line allows the ECB to lend dollars to European banks, creating dollar reserves in the process.[9] The New York Fed's assets increased by $7 billion, in the form of holdings of Euros in an account at the ECB; its liabilities also increased by $7 billion, in the form of increased deposits, corresponding to deposits in U.S. commercial banks held by the European banks that borrowed dollars from the ECB. The same day that the swap drawdown occurred, the $7 billion increase in foreign currency holdings was apportioned to all 12 Federal Reserve Banks according to the percentages listed in the fourth column of Table 4. The reduction in the New York Fed's foreign currency holdings and the increases in the other Reserve Banks' foreign currency holdings were balanced by increases in New York's ISA balance and decreases in the other Banks' ISA balances. Again, this example is summarized in T-account form for New York and Richmond, in Table 6.

[9] Data on swap line drawdowns are available at www.newyorkfed.org/markets/fxswap/fxswap_recent.cfm, and a detailed explanation of the swap facility is provided at www.federalreserve.gov/monetarypolicy/bst_liquidityswaps.htm.

Table 6 T-Accounts for Federal Reserve Banks, Foreign Currency Swap Example

	Federal Reserve Bank of New York	
	Assets	**Liabilities**
Step 1	+$7 billion Euros at ECB	+$7 billion commercial bank deposits (reserves)
Step 2	−$4.742 billion Euros at ECB	—
Step 3	+$4.742 billion ISA balances	—

	Federal Reserve Bank of Richmond	
	Assets	**Liabilities**
	+$1.448 billion Euros at ECB	—
	−$1.448 billion ISA balances	—

2. ALLOCATION OF SOMA TRANSACTIONS AND ANNUAL REBALANCING

Table 4 listed the percentages according to which foreign and domestic SOMA transactions were allocated to the 12 Reserve Banks in 2011 and 2012. These percentages are updated annually through a process that reflects ISA balances over the year and the composition across Districts of currency outstanding (for the domestic portfolio) and the composition across Districts of member bank capital (for the foreign portfolio). New York has by far the highest allocation percentage for both the foreign and domestic portfolios, but the percentages for the other 11 Banks varied widely in 2012, from a low of 0.41 percent of the foreign portfolio for the Minneapolis Fed, to a high of 20.69 percent of the foreign portfolio for the Richmond Fed. The remainder of this section describes how the percentages are determined. An important element of the domestic portfolio rebalancing is that it also involves an approximate "settling" of ISA balances. In contrast, the foreign portfolio rebalancing generates ISA transactions as an outcome, but they do not drive the process.

Domestic Portfolio

In April of each year, the 12 Reserve Banks' allocation percentages for the domestic SOMA portfolio are updated. We will use a hypothetical example for the Federal Reserve Bank of Richmond to explain how the process works. Before going into the details, we need to introduce the gold certificate account, an item on the asset side of each Federal Reserve Bank's balance sheet. The gold certificate account is a carryover from the time that the United States was on a gold standard.

Today, the Systemwide gold certificate account corresponds to the value of gold held by the U.S. Treasury. While the gold certificate account plays a role in the process described below, in no way do the Treasury's gold holdings restrict the quantity of currency or bank reserves that the Federal Reserve can issue.

1. Denote Richmond's average daily ISA balance for the preceding 12 months by B, and recall that we follow the H.4.1 release and put ISA on the asset side of the balance sheet. In the first step, the ISA balance is reduced by B, and there is an offsetting increase of B in the Richmond Bank's gold certificate account. If B is negative, then the ISA balance rises and the gold certificate account falls in this step.[10]

2. Denote the Systemwide ratio of the gold certificate account to the value of Federal Reserve notes by $\bar{\rho}$.[11] Denote the corresponding ratio for the Richmond Bank by ρ_R. In the second step, Richmond's gold certificate account is adjusted upward or downward—as appropriate—to equate the new ρ_R to $\bar{\rho}$. The offsetting balance sheet entry is a decrease or increase in Richmond's holdings of the domestic SOMA portfolio.

3. Denote the new ratio of Richmond's domestic SOMA portfolio holdings to the total domestic SOMA portfolio by δ. Until the following April, Richmond's allocation of the domestic SOMA portfolio will be given by δ.

The rebalancing process is undeniably complicated. However, some intuition can be gained by thinking about a hypothetical case where the allocation of securities purchases is always quickly accompanied by matching reserve flows. Each time the New York Fed purchases securities, an identical quantity of reserve liabilities is created, typically on the balance sheet of the New York Fed. A fraction of the securities are quickly allocated to the Richmond Fed. If reserves of the same magnitude then flow from the New York Fed to the Richmond Fed, there will be offsetting ISA transactions. If this occurs for every securities purchase, then in step 1 above there will be zero average ISA balance, and the only adjustment in April would occur because of different growth in Richmond Federal Reserve notes than in the System as a whole. This example makes clear that it is primarily the combination of

[10] As described in the Appendix, it is possible for the gold certificate account to become negative in step 1. But any negative balance would be reversed in step 2.

[11] The Systemwide value of the gold certificate account has not changed since 2006.

Table 7 Currency and Reserves by District (April 10, 2013), and SOMA Domestic Allocations for 2012

District	Currency (%)	Reserves (%)	SOMA (%)
Boston	3.33	1.64	2.43
New York	37.70	67.07	56.07
Philadelphia	3.31	2.01	3.31
Cleveland	4.34	1.10	2.54
Richmond	7.30	4.85	7.12
Atlanta	12.37	2.96	6.03
Chicago	6.75	3.72	5.55
St. Louis	2.61	0.79	1.56
Minneapolis	1.67	0.53	0.91
Kansas City	2.66	1.26	2.01
Dallas	6.96	2.71	3.89
San Francisco	11.01	11.35	8.60
System Total	100	100	100

differential growth in reserves and currency that leads to changes in a Reserve Bank's allocation percentage for the domestic portfolio.

To further illustrate the relationship between a Reserve Bank's share of liabilities and its allocation percentage, Table 7 lists each Bank's share of total reserves ("other deposits") and net Federal Reserve notes outstanding on April 10, 2013, together with the 2012 SOMA domestic allocation percentages from Table 4. For every Reserve Bank except San Francisco, the SOMA percentage lies between the Reserve Bank's share of currency and its share of reserves. As we have seen, in any given year the allocation is a complicated function of past history, ISA over the prior 12 months, and the distribution of Federal Reserve notes. However, the table shows that the distribution of currency (Federal Reserve notes) and reserves together are generally a good approximation to the SOMA allocation.

Finally, an important thing to note about the annual rebalancing process is that it generally does not result in a Reserve Bank's ISA balance moving to zero. This would only happen if the Reserve Bank's ISA balance on the day of rebalancing were equal to its average balance over the prior 12 months.

Foreign Portfolio

The annual foreign portfolio allocation percentages are determined in January, rather than April. As with the domestic portfolio, a one-time adjustment takes place to bring the account balances across Reserve Banks in line with the new percentages. However, whereas the

domestic allocations are determined by a complicated process involving prior-year ISA balances and the distribution of Federal Reserve notes, the foreign allocations derive in a simple way from the distribution of Reserve Bank capital. Each Reserve Bank has capital and surplus, based on the capital of the member banks in the respective Federal Reserve District (see Section 5 of the Federal Reserve Act for the details). Again, we will use a hypothetical example for the Federal Reserve Bank of Richmond to explain the annual process for determining the new foreign allocation and for reconciling the foreign portfolio.

Denote the Richmond Fed's share of the SOMA foreign portfolio by ϕ_0. Denote the Richmond Fed's share of Systemwide capital and surplus by κ. For the next year, changes to the SOMA foreign portfolio will be allocated to the Richmond Fed according to the ratio κ. There is also a one-time rebalancing, to equate Richmond's foreign portfolio share to its capital share. If the capital share is greater than the foreign portfolio share ($\kappa > \phi_0$), then the foreign portfolio is increased to make the new share, call it ϕ_1, equal to κ. And if $\kappa < \phi_0$, then Richmond's foreign portfolio balance is decreased so that $\phi_1 = \kappa$. In the former case, the increase in Richmond's foreign portfolio balance is offset by a decrease in Richmond's ISA balance. Likewise, if $\kappa < \phi_0$, there is an offsetting increase in Richmond's ISA balance. Effectively, Richmond is buying (or selling) shares in the SOMA foreign portfolio using its ISA balances.

Referring to columns 4 and 5 of Table 4, the differential allocation percentages among Reserve Banks for the foreign portfolio simply reflect different levels of capital of the member banks in each district. The Richmond Fed has a relatively large allocation percentage for the foreign portfolio because Bank of America, one of the four largest banks in the country, is a member bank located in the Richmond District.

If one is tracking Reserve Bank ISA balances, the annual adjustments in January and April are significant for two reasons. First, to the extent that there were persistently large ISA balances over the *prior* year, say because of significant changes in the size of the domestic SOMA portfolio, the April rebalancing would lead to large one-time ISA flows.[12] Second, to the extent there are significant changes in the size of the overall SOMA portfolio over the *coming* year, say because of asset purchases or sales, or swap line drawdowns, the new percentages will affect the ISA flows as the portfolio grows or shrinks.

[12] The foreign portfolio rebalancing in January would lead to large ISA flows if there were a sharp divergence between Reserve Banks' capital shares and their foreign portfolio shares. In order for this to happen, there would have had to be large changes in capital shares over the course of the year, presumably because of banking industry restructuring.

Comparison to TARGET2

The European Monetary Union has a similar character to the United States from a monetary perspective, in that it is composed of a system of central banks that together administer a single currency. Just as the ISA provides, and measures, a form of credit among Federal Reserve Banks, the TARGET2 system in Europe provides, and measures, a form of credit among the national central banks in Europe.[13] Because there is a wealth of literature describing how TARGET2 works in the Eurosystem, we will not go into any detail on that topic here, instead focusing on two important differences between TARGET2 and the ISA. One difference involves how the systems work, and it has received significant attention already.[14] The other difference involves the interpretation of TARGET2 versus ISA balances, which has received less attention.

A key operational difference between TARGET2 and the ISA involves rebalancing. In the Eurosystem, there is no regular administrative process corresponding to the Federal Reserve System's April ISA rebalancing. In principle then, it is possible for TARGET2 balances among countries in the European Monetary Union to grow arbitrarily large in absolute value. In practice, the European sovereign debt crisis was associated with persistently large positive TARGET2 balances for Germany, Netherlands, Luxembourg, and Finland, and persistently large negative TARGET2 balances for Ireland, Portugal, Greece, Spain, and Italy. However, since late 2012, the absolute level of TARGET2 balances has been declining in most of these countries.[15]

As we will see below, the ongoing increase in the Federal Reserve System's balance sheet, together with the limited tendency for reserve balances to flow from New York to the other Districts, means that without the annual rebalancing, New York—like the first group of European countries listed above—would have a persistently increasing ISA balance. Would such a scenario create the same uproar in the United States that it has created in Europe? Likely not, because (i) Federal Reserve Districts do not correspond to national, or even state borders, and (ii) the (hypothetical) accumulation of ISA balances in New York is associated with the fact that New York is a financial center, rather than with an especially strong economy in the New York Federal

[13] We say "a form of credit" because the national central banks and Federal Reserve Banks are only pseudo-independent of each other.

[14] See the references mentioned in the Introduction.

[15] For several of the national central banks, TARGET2 balances are easily accessible through the banks' official websites. The website www.eurocrisismonitor.com provides updated time series of all TARGET2 balances.

Reserve District. In fact, as Eichengreen, Mehl, and Chitu (2013) discuss, prior to 1975 annual rebalancing did not take place among Federal Reserve Banks. In principle, there was instead daily settlement across regional banks using gold certificates, but in practice "interdistrict accommodation operations" took place and balances did build up over time. Eichengreen, Mehl, and Chitu (2013, 4) argue that the build up of these balances "did not excite experts or the American public, nor in most cases did they trigger insurmountable tensions between regions."[16]

The second important difference between ISA and TARGET2 arises from the different degrees of financial integration within Europe and the United States. One element—albeit a relatively recent one—of the highly integrated U.S. financial system is the prominent role of interstate bank branching. Interstate bank branching and its corollary interdistrict bank branching mean that some bank deposits are located in a Federal Reserve District that is different than the one where the reserves backing that deposit are held. Because the location of reserves may not coincide with the residence of depositors, ISA flows may give misleading information about underlying financial flows.

Consider again the check clearing example from Table 1. Suppose JPMorgan Chase customers in Ohio write checks for $1 million to Bank of America customers in California. These transactions represent a transfer of bank deposits from residents of the Cleveland Federal Reserve District to residents of the San Francisco Federal Reserve District. However, JPMorgan Chase's reserve account is held with the Federal Reserve Bank of New York, and Bank of America's reserve account is held with the Federal Reserve Bank of Richmond. Based on ISA balances then, one would incorrectly interpret the transactions as representing a transfer of liquid assets from the New York District to the Richmond District.

In practice, the fraction of deposits with the property just described is quite large. For example, on June 30, 2013, JPMorgan Chase had customer deposits of $950 billion, but less than half of those deposits were held at branches within the New York Federal Reserve District. Or consider Bank of America, with customer deposits of $1.02 trillion, more than 45 percent of which were held in just four states *outside* the Richmond district: California ($241 billion), Florida ($81 billion), New York ($62 billion), and Texas ($82 billion).[17] These examples are much

[16] It should be noted as well that earlier (im)balances did tend to be driven by differential economic activity across regions, as opposed to FOMC-directed securities purchases or swap line drawdowns.

[17] The numbers in this paragraph are taken from the FDIC's Summary of Deposits website, www2.fdic.gov/sod/.

less prevalent in Europe: For the most part, transfers of deposits from a bank in Germany to a bank in Finland, for example, would represent transfers of deposits from German residents to Finnish residents.

3. INTERDISTRICT FLOWS DURING AND AFTER THE GREAT RECESSION

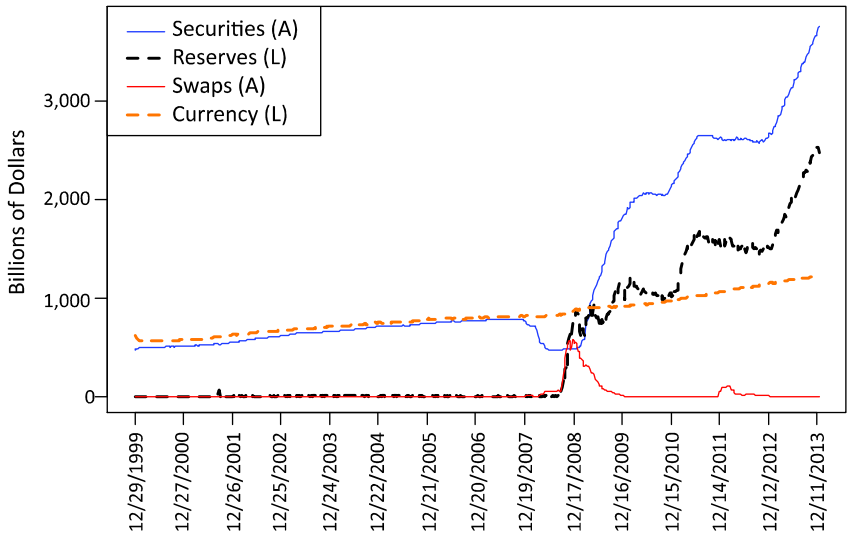
We turn now to actual data on ISA and other aspects of Reserve Bank balance sheets, concentrating on the post-2007 period. ISA behavior underwent a marked change after 2007 as a result of the Fed's credit programs, asset purchases, and swap lines with foreign central banks. After describing some of the more notable aspects of that behavior, we then suggest one way in which ISA behavior could provide useful information about the state of the economy as the Fed begins its exit from unconventional monetary policy.

Unconventional Monetary Policy and the ISA

Prior to September 2008, the balance sheets of the 12 Federal Reserve Banks grew at a fairly steady rate, mainly reflecting growth in currency demand as the economy grew. Secular growth does not necessarily imply changes in ISA balances, and both the volatility and absolute level of Reserve Bank ISA balances were low over this period. During the autumn of 2008, the Federal Reserve began paying interest on reserves at near market rates and lowered its Fed Funds rate target to near zero. Either one of these actions on its own would have severely reduced banks' incentive to economize on reserve holdings—previously a small fraction of currency outstanding. Simultaneously, and in a process that continues today, the Fed embarked on a series of credit expansion and asset purchase programs that dramatically increased the quantity of bank reserves: As of December 25, 2013, the aggregate level of reserves stood at $2.5 trillion, more than 239 times the level in early September 2008.[18] As described in Section 1, the asset purchases and central bank liquidity swaps that have generated much of this increase necessarily involve ISA transactions because the initial balance sheet increase at the New York Fed is subsequently allocated to the other 11 Banks. Thus, ISA balances at the 12 Reserve Banks behaved very differently after September 2008 than they had previously. In the remainder of

[18] See Keister and McAndrews (2009) and Ennis and Wolman (2012) for additional details on the behavior of bank reserves and the Federal Reserve System's balance sheet more generally.

Figure 1 Selected Components of Consolidated Federal Reserve Bank Balance Sheets

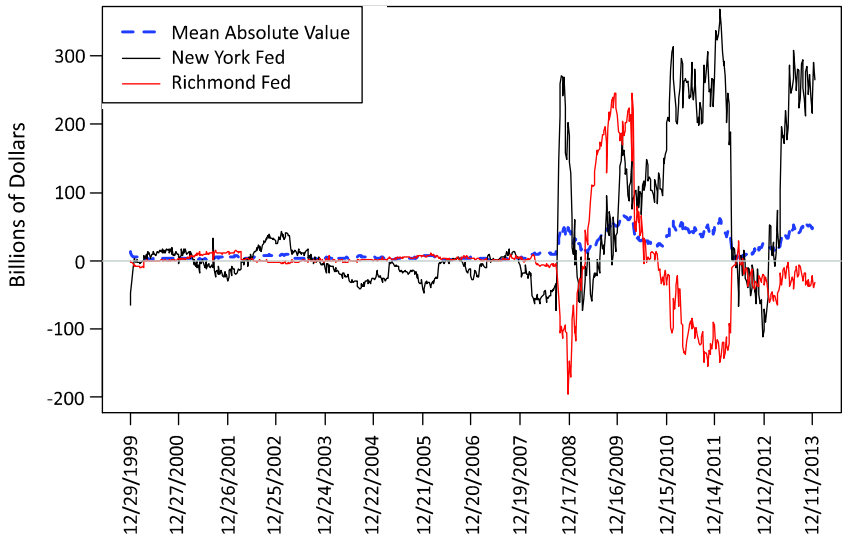


this section we discuss ISA behavior in the post-September 2008 period, concentrating on the Richmond and New York Banks.

Figure 1 displays four of the main components of the consolidated 12 Federal Reserve Bank balance sheets. Currency and reserves, which are liabilities to the Fed (hence denoted by an "L" in the legend), are plotted as the dashed orange and black lines, and the asset categories securities and swaps (hence "A" in the legend) are plotted as the solid blue and red lines. The figure reflects the discussion in the previous paragraph: In "normal times" securities grew steadily, hand in hand with currency. Once the large balance sheet expansions began in 2008, the dramatic increases in swaps and then securities were reflected in the growth of reserves, with currency remaining on a relatively stable upward trend.

For the same time period, Figure 2 plots ISA balances for the New York and Richmond Federal Reserve Banks, as well as the mean absolute value of ISA balances across all 12 Reserve Banks. There are several notable features of this figure. As stated above, before 2008, when currency and securities were growing steadily and reserves were

Figure 2 Interdistrict Settlement Account



low, both the level and volatility of ISA balances were low relative to their later behavior; this applies to Richmond, New York, and the entire System as reflected in the mean absolute value. That said, the swings in New York's ISA balance were large relative to the other Banks (compare the black line in Figure 2 to the red solid and blue dashed lines).

In a typical year before 2008, the New York Fed would be purchasing securities at a steady rate, and then immediately "selling" a significant fraction of those securities to the other 11 Banks, in exchange for ISA balances. This would tend to make New York's ISA balance increase over the course of the year ending in April, when the annual rebalancing of the domestic SOMA portfolio occurs. However, a close look at Figure 2 reveals that New York's ISA balance was just as likely to be decreasing over the year to April. The explanation may lie in the behavior of reserve balances: When the New York Fed purchases securities, the initial increase in reserves generally occurs in the accounts of banks in the New York District because the securities are sold by primary dealers, whose commercial bank accounts tend to be with New York banks. Over time, however (prior to 2008), the newly created

reserves would spread out across the System, roughly in proportion to economic activity, and be converted to currency. If the spreading out occurred before the conversion to currency, then it would involve an increase in ISA balances for other Banks and a decrease for New York, to offset New York's lower reserve account liabilities and other Banks' higher reserve account liabilities. Overall, ISA balances were low and stable at the other 11 Banks because, to a first approximation, the other 11 Banks were simply offsetting New York's fluctuations, with percentages similar to those in Table 4 (recall that the percentages are updated annually).

Beginning in September 2008, just as the size and composition of the consolidated Federal Reserve Banks' balance sheet began to change dramatically, so did the behavior of ISA balances. This occurred at the New York Fed as well as the other Reserve Banks. From the end of 1999 through September 10, 2008, the New York Fed's average absolute ISA balance was $17.1 billion; from September 17, 2008, through the end of 2013, New York's absolute ISA balance averaged $141.2 billion. For all Federal Reserve Banks, the corresponding increase was from $4.5 billion to $35.2 billion.[19]

While the entire post-2008 period has been characterized by high and volatile ISA balances, the behavior of New York and Richmond's ISA balances relative to the rest of the System divides into five distinct phases. In phase 1, from September 2008 through March 2009, New York's ISA balance rose and fell dramatically, and Richmond moved in opposite directions with somewhat smaller swings. Phase 1 is mainly accounted for by the behavior of swap lines. Swap line drawdowns increased from $62 billion on September 17, 2008, to their peak of $583 billion on December 10, and then by March 11, 2009, had fallen to $314 billion. As swap drawdowns rose and fell, New York's ISA balance would naturally rise and fall (Richmond's would fall and rise). In phase 2, roughly from March through the end of 2009, both New York and Richmond's ISA balances were increasing. For New York, this was due to the first round of LSAPs, and for Richmond it seems to have been due to an increase in deposits (reserves) that was quite large relative to other Banks (see Figure 4). Both Richmond and New York's ISA balances were relatively stable throughout 2010, apart from a large decrease for Richmond with the annual rebalancing of the domestic SOMA portfolio in April; because of Richmond's large average balance

[19] The calculation for all 12 Banks is as follows: First, calculate the weekly mean absolute balance across Banks, then average that balance across time to arrive at $4.5 billion and $35.2 billion for the two periods.

over the previous 12 months, the 2010 rebalancing involved reducing Richmond's ISA balance by approximately $175 billion.[20]

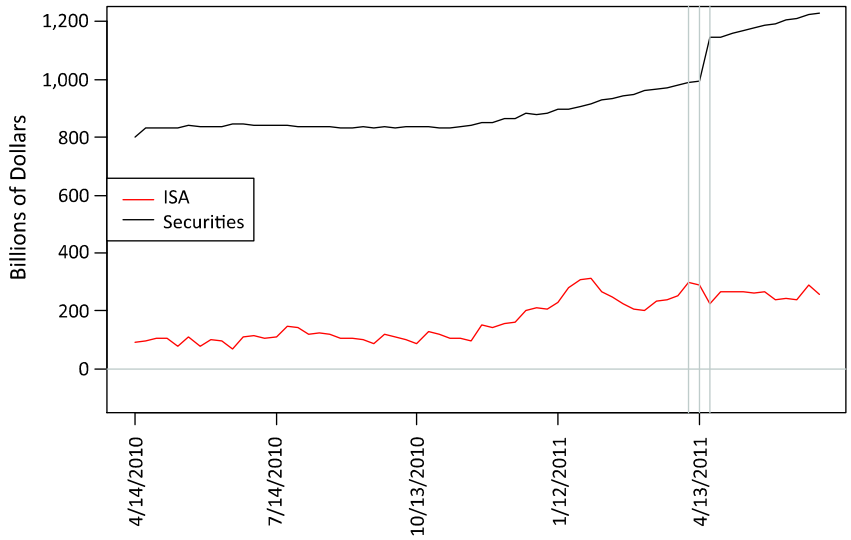
During phase 3, which lasted from late 2010 through the April 2012 domestic SOMA rebalancing, ISA balances in Richmond and New York were driven by the increase in reserves from the second LSAP program. The typical pattern associated with securities purchases occurred: New York's ISA balance increased as it allocated the newly purchased securities across the System, and Richmond's ISA balance decreased as it "purchased" securities from New York. These asset purchases ended in the middle of 2011, and ISA balances were relatively stable until the April 2012 rebalancing. At that time there was a large reallocation of securities from Richmond to New York, with a corresponding decrease in New York's ISA balance and an increase in Richmond's ISA balance; effectively, New York was purchasing back a similar but not identical quantity of securities from Richmond.

Regarding phase 3, there has been some speculation among commentators that rebalancing did not occur in April 2011. As evidence in favor of this view, Koning (2012) notes that while the New York Fed had an average ISA balance of around $147 billion over the previous 12 months, there is no evidence in the H.4.1 data of a similar-sized ISA decrease in April 2011. However, Koning also notes that the discrepancy may be a result of the inherent limitations in weekly data. In fact, this latter view is correct. Rebalancing did occur as usual, as can be confirmed by looking at the behavior of securities on the New York Fed's balance sheet.

Figure 3 zooms in on the behavior of the New York Fed's ISA and securities holdings, from April 2010 through June 2011. The three vertical lines in the figure represent April 6, April 13, and April 20, 2011. As described in Section 2, the annual domestic portfolio rebalancing for a Bank with positive ISA balance over the past year involves a decrease in its ISA balance and an equal-sized increase in its securities holding; the Bank is effectively purchasing securities with its ISA balance. Although New York's ISA did not display an unusual decrease in April 2011, its securities holdings did increase by $150 billion from April 13 to April 20. Securities were increasing steadily during that period because of the second LSAP program, but the rate of increase was nowhere close to $150 billion per week. The only plausible explanation for the $150 billion increase in securities is the annual rebalancing, which Koning indeed calculates ought to have been close to $150

[20] The number in the text is approximate because it is based on the weekly H.4 data, which incorporate all factors that affected the ISA during the week that settlement occurred.

Figure 3 New York: ISA and Securities around April 2011



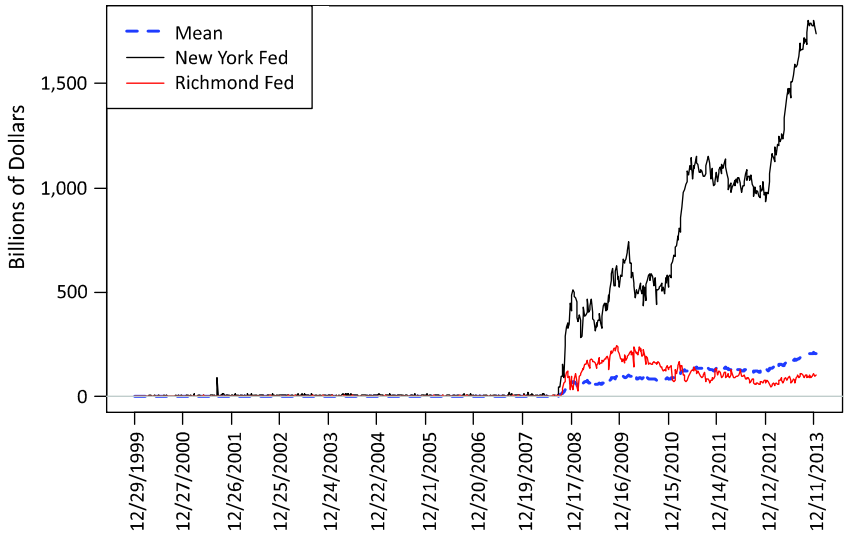
billion. The ISA change is not visible in the weekly data because it was partially offset by other factors unrelated to the rebalancing.

Phase 4, from April 2012 until late 2012, was characterized by declining ISA balances in both Richmond and New York. During this period, aggregate reserves were relatively stable (Figure 4), but deposit liabilities in both Richmond and New York were declining, with the offset coming from ISA balances. Evidently reserves were flowing out of Richmond and New York to the other Districts. Finally, phase 5 corresponds to the ongoing third LSAP program. New York's ISA balance has increased markedly from allocating the new securities purchases, and Richmond's balance has generally been declining since the last SOMA rebalancing in April 2013.

ISA Fluctuations as a Potential Signal for Monetary Policy

In comparing TARGET2 to ISA, we noted that the prevalence of interdistrict branching in the United States meant that ISA behavior was unable to provide the kind of information about cross-region

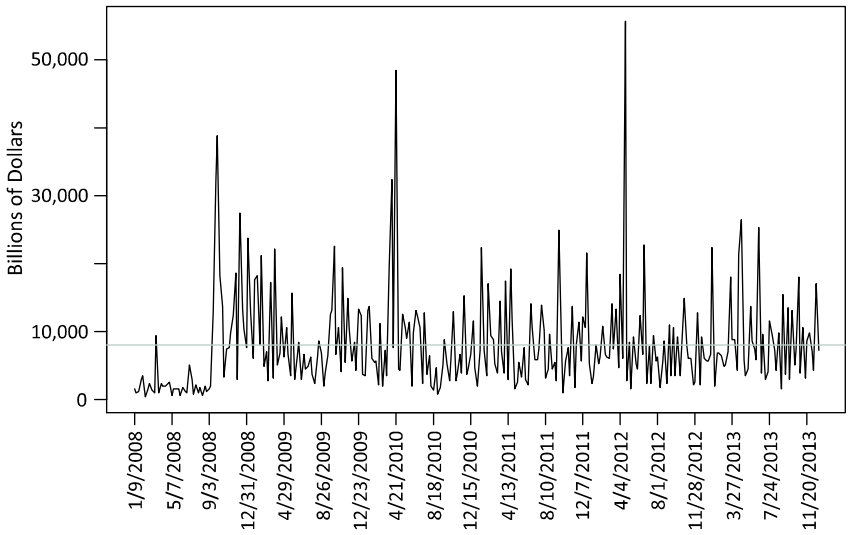
Figure 4 Deposits (Reserves)



payment flows that TARGET2 can provide. However, it should be clear from the example we used to make that point that ISA behavior does provide some information about payment flows across institutions. At the weekly level, only net flows across Federal Reserve Districts are captured, so flows across institutions within the same Federal Reserve District are missed entirely. Nonetheless, there may be some value in the information that is captured by ISA.

Starting in December 2013, the Federal Reserve began to reduce the pace of securities purchases in its third LSAP program. Assuming that the economic recovery continues, the tapering of asset purchases is likely to be the first stage in an exit from unconventional monetary policy, where the later stages will involve an increase in the federal funds rate target and a reduction in the Federal Reserve's securities holdings. Ennis and Wolman (2010, 2012) have argued that the large quantity of reserves outstanding makes it especially important that the Fed not fall behind the curve in raising its target for the federal funds rate. The financial flows represented by ISA fluctuations may provide one useful signal about the right time to raise that target.

Figure 5 Summary Statistic for Dispersion of ISA Changes



Informally, the idea is that if monetary policy were to fall behind the curve we would eventually expect to see inflation, but the inflation would likely be preceded by more rapid turnover of the monetary base (in particular, bank reserves). That increase in turnover would in turn be reflected in an increase in volatility of ISA balances. Figure 5 plots one measure of this volatility, from 2008 through 2013. For each Reserve Bank, we calculated the absolute valuable of the weekly change in the Bank's ISA balance, from the H.4.1 report. Then, for each week, we calculated the standard deviation of these changes across the 12 Banks. The jagged line in Figure 5 is the time series for this standard deviation, and the grey horizontal line is the mean over the period from January 2008 through December 2013. There are no surprises in Figure 5, given what we already know from the previous figures. In September 2009 there was a discrete upward shift in the dispersion measure, but since that time the series' behavior has been relatively steady, apart from spikes at the April rebalancing in 2010 and 2012. In the scenario where ISA behavior signals that it may be time for interest rates to rise, we would see an upward shift in the dispersion measure.

Anyone can track the dispersion measure in Figure 5, simply by downloading data from the Federal Reserve's website. As such, it may provide a useful way for the interested public to track monetary conditions. Policymakers themselves have access to the daily reserve balances of every financial institution with an account at a Federal Reserve Bank. They can therefore construct a more granular version of Figure 5, which begins with the absolute daily change in reserve balances for each account-holding institution, instead of the absolute weekly change in ISA balances for each Reserve Bank.

4. CONCLUSION

The massive expansion of the Federal Reserve System's balance sheet since 2008 has been accompanied by a notable increase in payment flows across Federal Reserve Districts. These payment flows are measured by the Federal Reserve's Interdistrict Settlement Account (ISA), much as fluctuations in TARGET2 balances measure payment flows across national central banks within the Eurosystem. There is, however, an important difference in the mechanics of the two systems; annual rebalancing occurs in the United States but not in Europe. In addition, because the U.S. banking system is highly integrated across regions, there are limits to the kind of information about payment flows that can be conveyed by ISA data.

Although the post-crisis period comprises several distinct phases of ISA behavior, as described in Section 3, the overall trend has been one in which the FOMC's asset purchase programs have tended to increase ISA balances (an asset) as well as deposit liabilities on the New York Fed's balance sheet. Absent the annual rebalancing process, described in Section 2, rough calculations suggest that New York's ISA balance would have risen to approximately $800 billion by the end of 2013, assuming that it started at zero at the beginning of 1999. Going forward however, as the asset purchase programs are eventually reversed, we should expect the behavior of ISA balances at New York and the other Banks to reverse as well. As long as the quantity of bank reserves remains large, the behavior of ISA balances may turn out to be a useful indicator of when the time has come for the fed funds target to rise.

APPENDIX: FORMAL DESCRIPTION OF ISA SETTLEMENT

What follows is a more formal statement of the process described in Section 2, for annual settlement of ISA using the domestic SOMA portfolio.

1. (a) Denote Richmond's average ISA balance for the preceding 12 months by B_R, and recall that we follow H.4.1 and put ISA on the asset side. In the first step, the ISA balance is reduced by B_R, and there is an offsetting increase of B_R in the Richmond Bank's asset item, "gold certificate account." If B_R is negative, then the ISA balance rises and the gold certificate account falls in this step.

   (b) Denote the Systemwide ratio of the gold certificate account to the value of Federal Reserve notes by $\bar{\rho}$. Denote the corresponding ratio for the Richmond Bank by ρ_R. In the second step, Richmond's gold certificate account is adjusted upward or downward—as appropriate—to equate the new ρ_R to $\bar{\rho}$. The offsetting balance sheet entry is a decrease or increase in Richmond's holdings of the domestic SOMA portfolio.

   (c) Denote the new ratio of Richmond's domestic SOMA portfolio holdings to the total domestic SOMA portfolio by δ. Until the following April, Richmond's allocation of the domestic SOMA portfolio will be given by δ.

   (d) Given

$I_{R,0} =$	Richmond's initial ISA balance
$B_R =$	Richmond's average ISA balance
$I_{R,1} =$	Richmond's new ISA balance
$G_{R,0} =$	Richmond's initial gold certificate account
$G_{R,1} =$	Richmond's "intermediate" gold certificate account
$G_{R,2} =$	Richmond's new gold certificate account
$G =$	System's gold certificate account
$N =$	System's Federal Reserve notes
$N_R =$	Richmond's Federal Reserve notes outstanding
$S_{R,0} =$	Richmond's initial SOMA holdings
$S_{R,1} =$	Richmond's new SOMA holdings
$S =$	System's SOMA holdings
$\delta_R =$	Richmond's new SOMA allocation percentage

   i. In step a, we have $I_{R,1} = I_{R,0} - B_R$ and $G_{R,1} = G_{R,0} + B_R$.

   ii. In step b, we have $G_{R,2} = G_{R,1} + \left(\frac{G}{N} N_R - G_{R,1}\right)$ and $S_{R,1} = S_{R,0} - \left(G_{R,2} - G_{R,1}\right)$.

   iii. Thus, for step c, $\delta_R = S_{R,1}/S$.

   iv. Note that $G_{R,1}$ is completely artificial. For an instant, a bank's gold certificate account could go highly negative

or could go higher than the System's total, though at every instance the total across Banks does sum to the System's total. We can rewrite the process without $G_{R,1}$ as $G_{R,2} = \frac{G}{N} N_R$ and $S_{R,1} = S_{R,0} - \left(\frac{G}{N} N_R - (G_{R,0} + B_R)\right)$. This makes it clear that Richmond's gold certificate account only changes to the extent that either (i) the System's ratio of gold certificate account to notes changes, or (ii) Richmond's notes quantity changes. And, Richmond's SOMA changes if (i) Richmond's gold certificate account changes, or (ii), more importantly in practice, if Richmond's ISA balance averaged something other than zero over the previous 12 months.

REFERENCES

Board of Governors of the Federal Reserve System. 2014. "Factors Affecting Reserve Balances. Federal Reserve Statistical Release H.4.1." Available at www.federalreserve.gov/releases/h41/.

Board of Governors of the Federal Reserve System. 2014. *Financial Accounting Manual for Federal Reserve Banks*. Available at www.federalreserve.gov/monetarypolicy/files/BSTfinaccountingmanual.pdf.

Cecchetti, Stephen G., Robert N. McCauley, and Patrick M. McGuire. 2012. "Interpreting TARGET2 balances." BIS Working Papers No. 393. Available at www.bis.org/publ/work393.pdf.

Eichengreen, Barry, Arnaud Mehl, and Livia Chitu. 2013. "Mutual Assistance between Federal Reserve Banks, 1913–1960 as Prolegomena to the TARGET2 Debate." Manuscript.

Ennis, Huberto M., and Alexander L. Wolman. 2010. "Excess Reserves and the New Challenges for Monetary Policy." Federal Reserve Bank of Richmond *Economic Brief* 10-03.

Ennis, Huberto M., and Alexander L. Wolman. 2012. "Large Excess Reserves in the U.S.: A View from the Cross-Section of Banks." Federal Reserve Bank of Richmond Working Paper No. 12-05.

Keister, Todd, and James McAndrews. 2009. "Why Are Banks Holding So Many Excess Reserves?" Federal Reserve Bank of New York *Current Issues in Economics and Finance* 15 (December).

Koning, J.P. 2012. "The Idiot's Guide to the Federal Reserve Interdistrict Settlement Account." Available at http://jpkoning.blogspot.ca/2012/02/idiots-guide-to-federal-reserve.html.

Lubik, Thomas A., and Karl Rhodes. 2012. "TARGET2: Symptom, Not Cause, of Eurozone Woes." Federal Reserve Bank of Richmond *Economic Brief* 12-08.

Whelan, Karl. 2012. "TARGET2 and Central Bank Balance Sheets." Available at www.karlwhelan.com/Papers/T2Paper-March2013.pdf.

Economic Quarterly—Volume 99, Number 2—Second Quarter 2013—Pages 143–162

Too Big to Manage? Two Book Reviews

Edward Simpson Prescott

Roddy Boyd's *Fatal Risk: A Cautionary Tale of AIG's Corporate Suicide* and Greg Farrell's *Crash of the Titans: Greed, Hubris and the Fall of Merrill Lynch and the Near Collapse of Bank of America* are interesting and informative books about two of the large financial firms that got into trouble and played an important role in the recent financial crisis. American International Group (AIG) was bailed out by the Federal Reserve and the federal government while Merrill Lynch almost certainly would have failed if it had not been acquired by Bank of America over that tumultuous weekend in which Lehman Brothers failed.

Both books cover, from the perspectives of these two firms, the events leading up to and during the financial panic of the autumn of 2008. The descriptions are useful and entertaining, but there are many other books on the financial crises that cover these events too. What these two books do provide that many other books do not is a window into how these two large firms were run, how they grew leading up to the crisis, and what decisions were made or not made that got the firms into trouble. What I want to do in this review is to use the books' analyses of AIG and Merrill Lynch to give some insight into how large financial institutions are run, their risks, why some of them failed in the recent crisis, and the implications for too-big-to-fail policy.[1]

The author would like to thank Arantxa Jarque, Sam Marshall, David Price, and John Weinberg for helpful comments. The views expressed in this article do not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: edward.prescott@rich.frb.org.

[1] Farrell's book covers much more than the buildup of risk that led to Merrill's near failure. It is also about John Thain's unsuccessful attempt to keep Merrill Lynch independent, its sale to Bank of America, and the ensuing after effects. It also discusses Bank of America, including some of its history. Because my interest in this review is limits on a person's ability to manage large financial firms, I won't discuss these parts of the book. Furthermore, in my mind, the history of Bank of America—and its

Both authors put the role of the CEO at the center of their stories. Boyd argues that AIG collapsed because the high-energy, aggressive Hank Greenberg built a firm with risks that his successor, Martin Sullivan, could not manage when he took over in 2005. Farrell argues that Merrill Lynch lost its independence because the ambitious, distant Stan O'Neal ripped up the old "Mother Merrill" culture when he took over in 2002. Like any good story, both books discuss the personalities of these leaders, their humble roots, how they interacted—or in some cases did not interact—with their subordinates, and how these character flaws contributed to the ending of their firms. These Shakespearean elements make for a good tragedy and, indeed, are essential to the story, but the focus on individuals runs the risk of hiding the real lesson of both books. In my view, both stories are ultimately about the limits of a leader's span of control, that is, the scope and scale of people and activities that a person can effectively manage. Both books provide evidence that these firms were so large, leveraged, and complicated that mistakes by leadership were fatal when the mortgage market declined. What is not addressed in either book, however, is the equally important lesson of why these two firms were able to grow to become so large, leveraged, and complicated in the first place. Later in this article, I will argue that the answer to that question lies in 40 years of federal policy of bailing out large financial firms.

1. AMERICAN INTERNATIONAL GROUP

When American International Group (AIG) was bailed out by the federal government in September 2008, it was a $1 trillion company with an astonishing reach. It operated worldwide, had a huge number of counterparties, and, in addition to supplying traditional insurance products like life insurance and property and casualty insurance, it leased aircraft, provided asset management services, sold annuities, insured stable value funds in pension plans, and was active in capital markets. It was involved in so many parts of the economy that it is not hard to see why it was viewed as too big to fail.

Boyd tells a convincing story about how AIG got to this point. He gives some background on the unusual history of AIG, but spends much of the book discussing Maurice "Hank" Greenberg, its CEO until

---

Queen City neighbor Wachovia—is really the story of the end of legal and regulatory restrictions on interstate and intrastate bank branching and the ensuing scramble among banks to be the "winner" in the acquisition game. For a book with more history of these two banks (as well as that of a third bank, the conservatively run "old" Wachovia), that gives some idea of why Charlotte of all places ended up as the second most important banking center in the United States, see Rick Rothacker's *Banktown*.

2005, and the growth of the financial products group, AIGFP. This unit issued the credit default swaps (CDS) that, along with losses in AIG's securities lending unit, were the main causes of AIG's collapse.

AIGFP was set up in 1987 as a joint venture with Howard Sosin, a former academic and a trader with Drexel Burnham Lambert. The vision of AIGFP was to use the AAA rating of AIG to fund derivative transactions, like interest rate swaps, at a lower cost than its competitors, and for most of its years, AIGFP seemed to do this very well.[2] Boyd describes what AIGFP did, but he spends a lot of time talking about its leaders. He describes Sosin's strong-willed personality and his conflicts with Greenberg. He also covers the succeeding years after Sosin was forced out in 1993, when AIGFP was first run by the calm Minnesotan Tom Savage and then, starting in 2001, by the hard charging, intimidating Joseph Cassano.

The AIGFP transactions that did so much damage to AIG were part of its CDS portfolio, and actually only a small portion it. A CDS is essentially an insurance contract written on the performance of some asset. AIGFP started providing CDS in 1998. These CDS were initially written on corporate debt, but over time AIGFP expanded the pool of assets it insured to include bank loans and, starting in 2004, collateralized debt obligations (CDO). A CDO is a security that receives cash from a trust that holds a bundle of loans, fixed-income securities, or other assets. From 2004 until the end of 2005, the CDOs that AIGFP insured included subprime mortgage-backed securities. Some of these CDS contained credit swap annexes that required AIG to post collateral if the value of the referenced security dropped in value. Downgrades to the referenced securities, as well as to AIG as a whole in 2008, required AIG to post large amounts of cash as collateral that it did not have in September 2008. The liquidity problems from these collateral calls and losses on its securities lending portfolio were the two most significant causes of its collapse.

The portion of AIGFP's CDS portfolio that caused so much trouble for AIG was, as mentioned earlier, proportionally small. As of September 2008, AIG insured about $360 billion of assets with CDS and only $55 billion of that was on CDOs that contained subprime mortgage securities (Congressional Oversight Panel 2010, 24). It was these latter CDOs that caused most of the losses and, furthermore, these losses came from just 125 of AIGFP's approximately 44,000 derivative contracts. Indeed, as profitable as AIGFP was, it was never that big a

[2] Using AIG's AAA rating to generate low-cost funding seems to have been a strategy of AIG's. That was one of the reasons, for example, that AIG bought the aircraft leasing business ILFC (Boyd 2011, 66).

percentage of AIG's income. For example, the Congressional Oversight Panel (2010, 23) reports that in 2006, AIGFP provided only about 7 percent of AIG's operating income.[3]

To understand how AIG got to the point where a relatively small portion of the firm could bring the rest of it down, it is necessary to understand something about AIG's history and the dominating role played in it by Hank Greenberg.

AIG was a very unusual company. It was founded by Cornelius Vander Starr in Shanghai in 1919. Starr ran the company until he appointed Greenberg as his successor in 1968. Under Greenberg, the company grew dramatically.[4] It expanded its insurance business and, in 1987, it entered capital markets through its joint venture with Sosin. It also had a very unusual long-term incentive scheme in which Greenberg would dole out shares of the Starr company—a byproduct of the corporate reorganization of AIG that he undertook when he first ran the company—that he also controlled, to loyal employees once they reached age 65. The promise of this long-term payout, which seems similar to the old investment banking partnership model, tied employees to AIG and gave them strong incentives to work hard and be loyal to the firm.

Boyd makes clear that much of the growth of AIG was due to the ambition and energy of Greenberg. The central role that Greenberg played in AIG is reflected in this description of Greenberg's management style (Boyd 2011, 132–3):

> All people who discuss Greenberg and his tenure at AIG eventually mention the beehive of his office. People came and went, orders were delivered—often in under one minute—and more people flow in and more orders are laid out....It was common for a division chief, earning well into seven figures, to be sitting in a chair next to the CFO as Greenberg sat behind his desk on the phone listening to someone from Tokyo while carrying on (possibly) related conversations with the division chief and CFO. Often, these conversations were truly material as to corporate strategy and direction.

The picture that one gets of AIG is that it was a somewhat decentralized organization, with an entrepreneurial culture, but in which there was effective corporate oversight in the form of Greenberg. Boyd gives a story about how Greenberg watched positions that, for as large a company as AIG, are relatively small (Boyd 2011, 75):

---

[3] The highest percentage it reached was about 12 percent in 2002.

[4] Greenberg resigned as CEO in 2005. This means that over an 86-year period, the firm only had two leaders.

> On many occasions, Davis and Rubin [two of the leaders of AIG Trading] had gotten called out of meetings, flagged down on vacations, interrupted in the middle of a big trade, and ordered to defend a certain position then on the books. The rub was that most every time it was the tail end of some big trade they were squaring away with a large customer and were in the process of selling. On a few occasions, they had made the mistake of attempting to reason with Greenberg, something to the effect of, "Hank, this is a $5 million position in yen futures/gold forwards/natural gas options. It's really liquid and pretty minimal in the scope of-"
>
> [Greenberg replying] "Hedge it, reinsure it, or there are consequences."

Boyd also writes (2011, 63):

> But any analysis of AIG's risk management begins and ends with Greenberg. Like a brilliant professor with a cluttered office, he knew where everything was and what it all meant. He was the risk management terminus, the ultimate arbiter of what was and was not acceptable. The problem was not that it didn't work, but that it worked so well. A generation of AIG employees learned to measure the risk they took so that it would be congruent with what Hank would tolerate. Investors and analysts happily assumed that a system like that would be in place forever more.
>
> It wouldn't be.

Boyd's view is that this dependence on Greenberg was AIG's weakness. If he were to leave and a lesser mortal stepped in his place, the system would break down, and this is what he argues happened when Martin Sullivan replaced Greenberg.

One of the most extraordinary things about how AIG lost Greenberg was the way it happened. Despite Greenberg turning 80 years old in 2005, he did not become ill or simply decide to retire. Instead, he was forced to leave because of the actions of Eliot Spitzer, the politically ambitious New York attorney general.

In the aftermath of the Enron accounting scandals, the political environment had shifted toward more aggressive enforcement of accounting violations. Based on several reinsurance transactions in 2000 that were of a questionable accounting nature, Spitzer went after AIG hard. Boyd's view is that Spitzer's legal case was not that strong, that he aggressively used leaks to the media to frame public opinion at the expense of the rule of law, and that he was driven by his political ambitions. Regardless of the merits of Spitzer's case, the end result was that Greenberg was forced out in February 2005 as head of AIG and replaced with Martin Sullivan. Partially because of the scandals, AIG

lost its coveted AAA rating. Furthermore, the attention of the Board of Directors and senior leadership was so focused on dealing with the legal risks from settlements with the attorney general and regulators that they were distracted from dealing with more traditional sources of risk.

It is after Greenberg left that AIGFP and Securities Lending made the decisions that got AIG into trouble. AIG kept writing CDS through the end of 2005 after Greenberg left in February of that year and even after AIG was downgraded from its AAA rating.[5] Furthermore, the increase in risk taken by the securities lending program started in the winter of 2005, also after Greenberg had left. AIG's securities lending program took the investment-grade securities that its various insurance subsidiaries owned and then lent them out for cash collateral. They then took this cash and, rather than lend it against safe securities like short-term Treasury securities, they lent it against risky securities such as subprime mortgage-backed securities.[6] Not only did the value of these securities drop, but they created liquidity risks because the cash lenders demanded their cash back and AIG was forced to sell these long-term securities precisely when mortgage markets were collapsing and becoming more illiquid.

Greenberg claims that once AIG was downgraded in early 2005, he would have stopped insuring the CDOs if he was still at the helm (Congressional Oversight Panel 2010, 27). Whether this is true is, of course, impossible to know. One distinction between the two regimes, however, is that under the Sullivan regime, there is evidence that AIG's senior management was unaware of the risks that AIGFP was actually taking. The collateral calls on AIG in the autumn of 2008 were based on contractual terms in annexes to the CDS contracts. Amazingly, corporate headquarters seems to have been unaware of the existence of these annexes (Boyd 2011, 325). This suggests that there were serious problems in AIG's controls and reporting systems. Indeed, the Congressional Oversight Panel (2010, 28) reports that AIG's auditor, PricewaterhouseCoopers, noted that in 2007 there were material weaknesses with the valuation of the CDS written by AIGFP on super senior CDO securities.

---

[5] Not all of the CDO risk can be attributed to these latter CDS. The CDOs that AIGFP insured had a feature called dynamic asset management (Congressional Oversight Panel 2010, 24), which means there is a collateral manager who replaces collateral as it is paid off according to the CDOs investment rules. Consequently, CDOs insured prior to Greenberg's departure would still have picked up some of the worst vintages of subprime loans that were made in 2006 and 2007.

[6] For example, in July 2007, one AIG unit discovered that 80 percent of its $540 million investment was really backed by mortgage-backed securities that could be considered subprime (Boyd 2011, 248).

Boyd points out some other weaknesses in AIG's information systems, such as the inability to get up-to-date financial information for AIG's units (Boyd 2011, 174), that suggest this was a more pervasive problem. The picture that one gets of AIG as a company is that the management information systems had some big weaknesses, but Greenberg's instincts and deep knowledge of the company compensated for these gaps. When Greenberg was forced to leave, this knowledge was lost and his replacement, Sullivan, was left with a company that was so big and complex that it had hidden risks.

2. MERRILL LYNCH

From 2002 to September 2007, E. Stanley O'Neal was the CEO of Merrill Lynch. He was hired in 1987 and quickly rose through the ranks. He became president in 2001 and acted decisively to first manage the operations of the firm after the terrorist attacks of September 11, 2001, and then to greatly reduce staff that was no longer needed because of the end of the tech boom. Partly because of his performance in this period, he was promoted to CEO in 2002, forcing out the previous CEO, David Komansky.

O'Neal greatly changed Merrill Lynch in both its strategic focus and its culture. Historically, the strength and focus of Merrill Lynch was its vast network of financial advisers—"the thundering herd"—who gave financial advice to Main Street America. Until O'Neal, Merrill's CEOs had been promoted from this line of business.[7] However, in the late 1990s, capital market and trading activities were growing relative to the financial advice business and were considered to be more promising. As CEO, O'Neal took this mandate and greatly expanded it.[8] The other dramatic change that O'Neal made to Merrill Lynch was to end its paternalistic culture of taking care of its employees. This culture gave the firm the nickname "Mother Merrill" and it meant, in practice, that mediocre performers were sometimes protected. When O'Neal reduced

[7] O'Neal actually ran wealth management for a short period of time before being promoted, but most of his career at Merrill was spent in other areas.

[8] For reporting purposes, Merrill broke its activities into two lines of business. The formal names of these businesses change over time, but one line of business consists mainly of wealth management and the other consists of capital market activities, like trading, as well as investment banking services, e.g., merger and acquisition advice. In 1998, the wealth management business had net revenues of $11.3 billion while the trading/investment banking line of business had net revenues of only $6.5 billion. By 2006, the proportional importance of the two units had almost reversed. The wealth management business had net revenues of $12.1 billion, while the trading/investment banking business, which includes the fixed-income, commodities, and currencies unit discussed later, had net revenues of $18.9 billion. (Source: Merrill Lynch Annual Reports 1998, 2006.)

staff, he did so dramatically by laying off 22,000 people, or nearly 30 percent of the firm's employees.

Farrell's view is that in destroying this old culture, which he thinks did need to be replaced or at least altered, O'Neal destroyed some of the important checks on risk-taking that had existed at the firm. First, paternalistic cultures tend to be more risk averse. Second, as part of the layoffs, he eliminated many executives who were associated with the old regime or were a potential threat to him and replaced them with a younger, more diverse group that was loyal to him (Farrell 2010, 89). What arose in its place was a culture missing strong independent executives willing to challenge O'Neal on decisions. Farrell believes it was these conditions that allowed for the decisions that caused Merrill Lynch's problems.

Merrill Lynch's biggest problems came from its fixed-income, commodities, and currencies, or FICC, line of business. One part of this business was to underwrite, or create, CDOs. A CDO underwriter buys the fixed-income securities that go into the CDO and structures the securities.

By 2004, Merrill Lynch was the largest underwriter of CDOs (Barnett-Hart 2009). As the housing boom grew, the volume of CDOs grew and many of them included mortgages, particularly subprime ones. Like many of the other investment banks, Merrill Lynch bought a subprime originator, First Franklin, in 2006 to vertically integrate the supply of mortgages.

A significant risk for a CDO underwriting firm is that it will not be able to sell all the CDOs it creates or, if it can't even put the CDO together, the assets that it bought in the first place. This was what happened to Merrill Lynch. In late 2006 and the first half of 2007, as most everyone else was getting out of this business, they kept underwriting CDOs. In the first half of 2007, Merrill underwrote $34 billion in CDOs, most of which ended up on its balance sheet because investors had stopped buying them (Farrell 2010, 18). Furthermore, these CDOs were backed by particularly risky collateral, namely, subprime loans made at the peak of the boom as well as risky tranches from other CDOs.

It was these positions that contributed the most to Merrill's troubles. At the end of the second quarter of 2007, before the write downs started, Merrill Lynch had a balance sheet of slightly over $1 trillion, and, like the other investment banks, it was highly leveraged, so it only had equity capital of $42 billion.[9] Amazingly, these CDO holdings

[9] Source: Merrill Lynch 10-Q, second quarter 2007.

performed so poorly that they lost most of their value over the next year, which wiped out much of this capital. (Merrill did raise capital over this period and had earnings in some other parts of the firm, which offset some of these losses.)[10] As Farrell (2010, 34) puts it,

> Merrill Lynch had just violated the cardinal rule of every financial institution on Wall Street, which holds that no one business unit should ever be given enough leeway to sink the entire firm.

To put this in perspective, in 2006 FICC's revenue net of interest expense was about $7.5 billion, which was about 22 percent of Merrill's total revenue (Merrill Lynch 2007 annual report). Furthermore, FICC not only underwrote CDOs, but also traded in currencies, commodities, and other fixed-income securities, so the 22 percent upper bound is probably far from the actual amount.

Farrell ties this disastrous buildup in risk to the hiring decisions made by O'Neal and one of his chief lieutenants, Ahmass Fakahany. In 2006, when FICC was created as a separate unit within the trading group, the head of sales and trading, Dow Kim, had to decide who would head it. Kim's first choice was an internal candidate named Jeff Kronthal who had experience with mortgage-backed securities, understood risk, and had been at Merrill Lynch since 1989. Furthermore, he had recently become cautious about the real estate market (Farrell 2010, 24). Kim's second choice was Jack DiMaio, an outsider, who had run a hedge fund and, as a consequence of that experience, understood risk. Unfortunately, neither O'Neal nor Fakahany (to whom Kim reported) wanted Kronthal or DiMaio. Instead, they wanted Osman Semerci, whom they had pegged as a rising star at Merrill. Kim was reluctant to hire him because of his lack of experience in risk but did what his bosses wanted (Farrell 2010, 25).

Semerci's background was in sales. He started in Merrill in retail and moved to institutional sales and did very well at that. However, he did not have much experience with risk and Farrell describes his promotion, with some hyperbole, as "[taking a] salesman with the instinct of a riverboat gambler and making him general manager of the casino" (Farrell 2010, 25). One month after Semerci took over in July 2006, Kronthal, along with a group of experienced traders, was fired.

---

[10] Merrill's 2007 and 2008 10-Ks give more details on FICC's losses. Over this two-year period, they wrote down their CDOs by $26.9 billion, wrote down U.S. subprime mortgages by $14.0 billion, adjusted the value of their hedges down by $13.0 billion, and wrote down subprime securities by $7.2 billion. The total was $61.1 billion over this two-year period.

Ostensibly, the buildup of Merrill's CDO exposure was due to a bad hiring decision, but this would not be the first time a corporate CEO hired the wrong person for a job. What is particularly troubling is that outside of FICC, the rest of Merrill seemed unaware of the size of the CDO position. Farrell does not provide the details on what the risk management, accounting, and other control functions in the firm were measuring with respect to the CDOs, but several stories he reports suggest that these systems were lacking.

Particularly illuminating were the difficulties that several high-up executives faced in determining just how much CDO exposure FICC built up under Semerci. At a July 2007 board meeting, Laurence Tosi, who was the chief operating officer of global markets and investment banking (which FICC was part of), learned that FICC had accumulated $31 billion of CDOs on its balance sheet, yet claimed they had minimal mortgage exposure. He was skeptical (Farrell 2010, 17–18) and tried to figure out just how much risk FICC really had (Farrell 2010, 16).[11] Furthermore, at about the same time a former risk executive named John Breit started his own attempt to figure out the true exposure after hearing about it from some junior quantitative analysts at a conference. Farrell describes the difficulties they faced in tracking down the exposures, mainly because the information was tightly controlled by Semerci and his staff was afraid of talk to non-FICC staff about these matters.

Farrell puts the positions that Semerci built up as the proximate cause of Merrill's failure and he believes that O'Neal did not realize how much CDO exposure was building up.[12] Nevertheless, he blames O'Neal and Fakahany for Merrill's troubles because they pushed for Semerci's promotion and, more importantly, O'Neal fostered a culture that eviscerated some of the checks that existed under the old Mother Merrill culture, which might have prevented Semerci's promotion and him from building up the large CDO exposure.

---

[11] Some of FICC's risk would not have shown up in accounting numbers because it was hedged. In order to sell AAA CDO securities, Merrill had traditionally bought protection from AIGFP that made the securities more appealing to investors. However, AIG stopped providing this service on subprime-backed securities in late 2005. Consequently, by 2007 Merrill was holding on to the AAA portions and hedged them by buying insurance from the monoline insurers. The monoline insurers were pretty thinly capitalized, and, given the nature of their business, couldn't really provide much insurance against big aggregate shocks, so these hedges were not that useful and later were written down in value.

[12] McLean and Nocera (2010) also believe that O'Neal was unaware of the size of the exposure. Furthermore, they think Kim, who left Merrill in May 2007, was unaware of it as well (McLean and Nocera 2010, 314).

3. SPAN OF CONTROL AND TOO BIG TO MANAGE

Despite AIG being primarily an insurance company and Merrill Lynch being primarily an investment bank, they had several features in common. First, both were very large and complex. At the end of 2006, AIG had $979 billion in assets, and Merrill had $841 billion in assets. Second, both were highly leveraged. At the end of 2006, AIG's leverage ratio was nearly 10, while Merrill's was nearly 22. Third, both got into trouble mainly from the actions of one or two units within their firm. Fourth, and this is the major thesis of the authors, neither firm's CEO had a good system in place for preventing the buildup of risk, or even recognizing it, in portions of their firm.

In the span of control model used in economics to study the size of firms (e.g., Lucas [1978]), managers differ in their ability to manage people and other resources. The more capable the manager is, the more people and activities he can effectively manage. If the market allocates resources to managers efficiently, then each manager or CEO of a firm gets the right amount of inputs. But if for some reason the market does not do this efficiently, then the CEO and his management team get the wrong amount.[13]

What seemed to happen in the case of Merrill and AIG is that they got too much capital and became too large to effectively manage. In AIG's case, the "system" for controlling risk was so dependent on Hank Greenberg that when he was forced to leave, it stopped working. His successors were left with a very large, complex organization in which a proportionally small but complex part was able to take enough risk to sink the organization. Similarly, while the Merrill collapse looks to be due to a bad hiring decision, it should not be forgotten that Merrill was a $1 trillion firm, and there was a lot more going on than just the CDO underwriting activities of FICC. For a firm of that size, a $30 billion exposure is a relatively small percentage of the balance sheet. The fatal mistake was to develop a corporate culture that did not recognize how risky that line of business could be and then allowing a risk-taker to run it.

There are other examples where the failure of one small part of a financial firm caused it to fail. One such famous case was the failure of Barings Bank in 1995. Barings failed because a single trader named Nick Leeson was able to use his control over back office functions to hide enormous bets that he took on the Japanese and Singaporean exchanges—bets that ultimately failed (Kuprianov 1995).

[13] The Appendix contains a span of control model where too-big-to-fail policies lead financial firms to become inefficiently big.

The lack of a proper control environment at Barings is an example of a management failure, though by recent standards Barings was neither a particularly large nor a particularly complicated firm. However, one bank whose troubles can be tied to growing too large was UBS. UBS made a strategic decision to expand its fixed-income business in 2005 near the end of the mortgage boom. However, as the UBS Shareholder's report (2008) documents, pricing of internal funding encouraged the accumulation of AAA-related CDO positions. One division would originate these CDOs and another division would buy them. Risk measurement did not fully pick up exposures, partly because they relied on the ratings, but also because information systems reported net (inclusive of hedges that turned out to be too small or not very good) rather than gross exposures.[14] The report concludes that senior management did not intend to take a lot of risk, but instead were unaware of how much risk the bank was really exposed to. Partly because of these losses, UBS was later bailed out by the Swiss National Bank.

Where the two books have a limitation is that there is a lack of detail about the risk management and other information systems used by the two companies. There are bits and pieces of evidence that suggest neither firm's systems were up to the task, but what could really cement this conclusion would be an in-depth analysis by someone with unfettered access to insiders and management reporting systems, like was done by UBS.[15] Then we would have a better sense of how much of the risk that was taken was due to inadequate measurement systems, how much was due to conscious risk-taking, and how much was just bad luck. Both authors had to work with what they could determine from public sources, as well as whoever was willing to talk with them, often off the record, so this criticism is not directed at them.

While these weaknesses in internal risk management and management information systems are important to investigate, it needs to be recognized that any system will eventually fail. What the AIG, Merrill Lynch, and UBS cases demonstrate is that diversification does not

[14] The poor quality of internal information seems to have been a problem at numerous large financial firms during this crisis. Kirsten Grind's book *The Lost Bank: The Story of Washington Mutual* details the rise and fall of this huge West Coast thrift. She reports that in its rapid accumulation of other banks and thrifts, Washington Mutual, by 2004, ended up with 12 different mortgage information systems and did not consolidate them, partially because its mortgage business was doing so well (Grind 2012, 99). Furthermore, when the market started to turn, the lack of attention to integrating data systems made it difficult for Washington Mutual to track the characteristics of its mortgage portfolio (Grind 2012, 165). So much for technological economies of scale in banking!

[15] The Congressional Oversight Panel (2010) report has some information along these lines for AIG.

always reduce risk for a financial firm. As the scope of a firm's activities grow, these activities become harder to evaluate and control. If losses from a particular activity can be large enough to sink the firm and the other activities of the firm can't function on their own, then failure of a single part of a firm can be disastrous. For financial firms that are highly leveraged and dependent on short-term debt, mistakes by management make this possibility even more likely. If a firm is involved in too many activities, then more diversification is really less.

4. TOO BIG TO FAIL AND TOO BIG TO MANAGE

So what might have led these two firms (and others) to get so large and complicated? Why might they have grown to exceed their managers' span of control? Some of it was certainly the housing boom. Most financial institutions did well in this period, so it was easy to grow. Nevertheless, another important factor at work, which neither author discusses, is that both firms were large enough that they could reasonably be considered to be too big to fail. This meant that their creditors could monitor them less carefully and charge less to lend to them. As a consequence, both firms could get larger and more complex than they would have otherwise. Indeed, Greenberg's strategy was to use the funding advantage that came with AIG's AAA rating to fund AIGFP's positions at a lower cost than its competitors, and that is one reason this unit, and others, could enter into so many transactions and grow.

The defining characteristic of U.S. financial regulatory actions over the last 40 years has been to intervene to prevent failures of large financial firms and to bail out short-term creditors of banks. The origins of this policy can be found in Sprague (1986), who describes a succession of bailouts made by the Federal Deposit Insurance Corporation (FDIC) from the early 1970s through the mid-1980s.[16] The first large one was Bank of Commonwealth, a $1.2 billion bank in Detroit. The next large one was of First Pennsylvania in 1980, a $9 billion bank that made a disastrous interest rate bet.[17] Finally, in 1984 there was the big bailout at the time, Continental Illinois, which is when the term "too big to fail" spread widely in public discourse.

[16] For an excellent book on too big to fail, see Stern and Feldman (2009).

[17] A large bank that was almost bailed out in 1983 was Seafirst, a $9 billion bank in Seattle that was heavily exposed to Penn Square, a bank that failed in 1982. Sprague (1986) reports that a $250 million loan from the FDIC was prepared and ready to be made in case Seafirst could not find a buyer. Fortunately for the FDIC, Bank of America bought the bank at the last minute.

Continental Illinois was a $30 billion bank that was mainly funded by uninsured deposits in the wholesale market. Furthermore, it had an extensive network of correspondents and counterparties. When Continental Illinois got into trouble, its wholesale lenders started pulling their money out. Bank regulators were so worried about the contagion effects of its failure that the Federal Reserve made extensive discount window loans that allowed uninsured depositors to withdraw their money and the FDIC took partial ownership.

As Hetzel (1991, 2012) documents, Continental Illinois was not the only bank for which Federal Reserve discount window lending was used to prevent a sudden failure. It was also used in the periods leading up to the failures of Franklin National in 1974 and the National Bank of Washington in 1990 and, in both cases, the emergency lending gave uninsured depositors time to get much of their money out of the bank before it failed. While there are exceptions, in general uninsured depositors rarely lose money in a bank failure.

While any doubts about whether nonbank financial firms like AIG or Merrill Lynch were too big to fail were erased by the financial crisis, what did creditors think before the crisis when these firms were growing? Did they think that they would they receive the same treatment as a bank in trouble? Based on the precedents discussed above, there are good reasons to think that they would have. Merrill Lynch funded its holdings of mortgage-backed securities by using short-term repo markets, which are essentially short-term loans and a bit like deposits. Failure in the repo market would be very disruptive. AIG's credit default swaps were held by many counterparties and some of them might have failed if AIG had failed, much like many correspondents and other banks might have failed if Continental Illinois had failed. Finally, there are precedents for financial regulators to intervene at nonbank financial firms and in financial markets. For example, when the hedge fund Long-Term Capital Management failed in 1998, the New York Fed put its creditors together—mainly the large commercial banks and investment banks—so that they would agree to put capital into the fund and avoid rapidly liquidating its assets. In 1987, when the stock market dramatically dropped, many broker-dealers were close to failing, but regulators pressured banks to lend to them to keep them functioning.[18]

It is well recognized that the safety net can encourage risk-taking, as in the infamous "gambling for resurrection" that some of the savings

[18] An extremely high fraction of financial liabilities are explicitly or implicitly backed by the federal government. Marshall, Pellerin, and Walter (2013) estimate that, as of the end of 2011, 57 percent of financial liabilities in the United States are explicitly or implicitly backed by the federal government.

and loans engaged in during the 1980s (see White [1991]). Sprague's description of how both Bank of Commonwealth and First Pennsylvania bought long-term securities, betting that interest rates would fall (but instead rose), seems to fit this description (Sprague 1986, 86).

But there is a second, indirect way in which the safety net encourages risk. Both AIG and Merrill Lynch seemed to have gotten too big and complicated for what their management could handle.[19] Now, in a sense, these two mechanisms are one and the same. After all, consciously becoming large and complicated is a way to become riskier, but knowingly taking a risky bet seems to have some differences from stumbling into a risky bet. My reading of the books is that both authors believe that the CEOs were unaware of just how much risk their firms were exposed to. They were too removed from the activities on the ground to understand the risks, while the enormous profits of the mortgage boom years masked some of the signals that might have warned them earlier about what was really going on.

5. CONCLUSION

Both books contain many other interesting insights into AIG, Merrill Lynch, other firms, and financial markets. Boyd's description of the history of AIG, with its international origins and Greenberg's connections to world leaders, makes one wonder about the political economy of the insurance business, while AIG's use of shares in Starr as a long-term incentive is worth knowing more about, particularly with the move in bank regulation toward pushing banks to use more deferred compensation. Similarly, Farrell describes the unusually powerful role played at Bank of America by its human resources department and, as a former financial reporter (he used to work for the *Financial Times*), he has special insight into how information makes its way to the public. For example, he makes it quite clear that executives at large financial firms are just as willing as Washington officials to strategically leak information to reporters.

While reading the books, the emphasis on Greenberg and O'Neal makes it tempting to look at the failure of both firms solely as failures of their CEOs. But behind both stories are really two important themes that transcend any individual. The first is that 40 years of bailing out financial firms and short-term creditors led us to the point where some financial firms are encouraged to get too leveraged, too complex, and

[19] For a description of the traditional risk-shifting model used to study bank risk-taking, see Prescott (2001). For a simple model of an alternative way in which the safety net increases risk, and which is along the lines of this review, see the Appendix.

too big for their own, or anyone else's, good. The second theme is that there are plenty of financial activities that can develop large exposures to risk and, when one of these fails, the losses can be so large that they are catastrophic and bring down the rest of the firm.

Where the two books excel is that they demonstrate how dangerous a bad decision can be in a large, leveraged, complex financial firm. A managerial mistake, either intentional or unintentional, can bring down a financial firm. If the firm is small, then such a mistake will likely cause failure, but the consequences won't be that severe. Put all of these activities into one firm and the same mistake will be less likely to cause a failure, but if a big enough mistake happens, the consequences will be a whole lot worse.

APPENDIX

This appendix works through a basic span of control model that formalizes the idea expressed in this review that large financial firms can get so large that they are riskier than is socially optimal. The model is a version of Lucas (1978) in which managers of varying talent levels manage capital.[20] The model can be used to characterize industries in which the size distribution is skewed to the right, that is, there are a few large firms and lots of small firms, which is the pattern in many industries and increasingly so in financial intermediation. The better a manager is, the more capital he manages. However, we add government bailouts that lower the cost of capital to large banks. As a consequence, the most talented managers manage a bigger bank than is socially optimal.

There is a cumulative distribution function, $H(t)$, of individuals with managerial talent t. An individual may either be a manager or a worker. A manager rents capital, k, and tries to produce output.[21] A manager is successful with probability $f(t,k)$. If successful, he produces $tg(k)$, and if he is not successful, he produces zero. We assume that $g(k)$ is increasing and concave in k and that $f(t,k)$ is linear and decreasing

[20] The model is also related to Ennis and Malek (2005), who develop a model of a large number of ex ante identical banks, each of which chooses its size and risk. Deposit insurance and too-big-to-fail policies encourage each bank to get inefficiently large and take on an inefficiently high amount of risk.

[21] Capital here is simply the funds invested in the firm. To keep the model simple, all the invested funds are treated like debt.

in k. The linearity is a strong assumption, but greatly facilitates the analysis. We also assume that $f(t,k)g(k)$ is increasing and concave in k for the range of capital relevant for this problem. This assumption ensures that the banks are in the region of capital where getting bigger still increases expected revenue.

The rental rate on capital equals its expected return; its risk-free rental rate is r. If an individual becomes a worker, his income is w. Both w and r are exogenous. Finally, each individual maximizes his expected income.

We model too-big-to-fail banks by assuming that if one of these banks fails, the owners of its capital are repaid their principal and still receive their interest. For simplicity, we assume that all banks with managerial talent $t \geq t^b$ are too big to fail, which means they only have to pay out the risk-free rate, r, when they are successful.[22] We also assume that all people with talent $t \geq t^b$ find it worthwhile to be managers; this way there will be banks that are not too big to fail and others that are. The decision for too-big-to-fail managers is how much capital to rent. They solve

$$\max_k f(t,k)(tg(k) - rk).$$

A linear equation times a concave function is concave, so this equation is concave and the first-order condition is necessary and sufficient for characterizing an optimum. It is

$$f_2(t,k)tg(k) + f(t,k)tg'(k) = f_2(t,k)rk + f(t,k)r. \tag{1}$$

For a manager who is not too big to fail, that is, $t < t^b$, the interest rate that he pays is $r/f(t,k)$, which reflects the probability that the owners of the capital might not get it back. His objective function is

$$\max_k f(t,k)tg(k) - rk.$$

The first-order condition is

$$f_2(t,k)tg(k) + f(t,k)tg'(k) = r. \tag{2}$$

Let $k^*(t)$ be the optimal amount of rental capital for a $t < t^b$ individual. A person with this level of talent will be a manager if

$$f(t,k^*(t))tg(k^*(t)) - rk^*(t) \geq w.$$

It is straightforward to show that $\frac{\partial k^*(t)}{\partial t} > 0$ and that a manager's profits are increasing in t. Therefore, there is a marginal manager, t^z,

[22] The more natural alternative is to make the too-big-to-fail cutoff depend on the amount of capital a bank manages, but that complicates the analysis because it creates a discrete choice for some banks of whether to exceed the too-big-to-fail threshold.

who is indifferent between being a worker and a manager. People sort into jobs according to the following rule

$t < t^z$ → workers
$t^z \leq t < t^b$ → manages a bank that can fail
$t \geq t^b$ → manages a too-big-to-fail bank.

The distortion in this economy is that capital for too-big-to-fail banks is subsidized. Not surprisingly, this means that these banks get inefficiently large. To see this, compare (1) with (2) for a fixed level of k. The former equation characterizes the amount of capital chosen by a bank with the too-big-to-fail subsidy, and the second equation characterizes the capital without the subsidy. The left-hand side of these two equations are identical and, by assumption, decreasing in k. Furthermore, comparing the right-hand sides of these two equations, observe that $f_2(t,k)rk + f(t,k)r < r$. Consequently, a k that satisfies (1) is more than a k that satisfies (2).

Too-big-to-fail banks are inefficiently big, and they fail more often than they would without the subsidy. Interestingly, in the debate about the quantitative effects of too big to fail, the spread in interest rates of bonds between the largest banks and small (but still large) banks is sometimes used to measure the size of the subsidy. In this model, this spread does not measure the subsidy, since the subsidy is the difference in the interest rate that would have been paid by the too-big-to-fail bank if it could fail and the risk-free rate. Furthermore, in the absence of the subsidy, the too-big-to-fail bank would be smaller, fail less frequently, and be more productive. Measuring the interest spread does not measure these effects either.

Decisions by managers with $t^z \leq t < t^b$ are not affected by the subsidy. Neither the size of a non-too-big-to-fail bank is affected nor who is the marginal manager because r and w are exogenous. This would not be true if r and w were endogenous.[23]

In this model, one solution to the distortion is a tax on firm size, or in a more general model, a tax on the insured liabilities of the too-big-to-fail banks. One proposal discussed in policy circles for getting rid of too big to fail is to cap bank size. In this model, that would mean capping the banks to the size corresponding to the largest non-too-big-to-fail bank. This would, of course, eliminate too big to fail, but as this model makes clear, it would do so at a cost, possibly a substantial one. In particular, the most productive banks—the ones run by the high

[23] In all likelihood, the general equilibrium effects need not be trivial. The subsidized capital moves capital through the banking system, which could lead to overinvestment. This in turn would affect the capital-labor ratio and, thus, r and w.

talent managers—would be artificially small, thus reducing banking sector productivity.

REFERENCES

Barnett-Hart, Anna Katherine. 2009. *The Story of the CDO Market Meltdown: An Empirical Analysis.* Cambridge, Mass.: Harvard University.

Boyd, Roddy. 2011. *Fatal Risk: A Cautionary Tale of AIG's Corporate Suicide.* Hoboken, N.J.: John Wiley & Sons, Inc.

Congressional Oversight Panel. 2010. "The AIG Rescue, Its Impact on Markets, and the Government's Exit Strategy." June Oversight Report. Washington, D.C.: U.S. Government Printing Office (June 10).

Ennis, Huberto M., and H. S. Malek. 2005. "Bank Risk of Failure and the Too-Big-to-Fail Policy." Federal Reserve Bank of Richmond *Economic Quarterly* 91 (Spring): 21–44.

Farrell, Greg. 2010. *Crash of the Titans: Greed, Hubris, the Fall of Merrill Lynch, and the Near-Collapse of Bank of America.* New York: Crown Business.

Grind, Kristen. 2012. *The Lost Bank: The Story of Washington Mutual, The Biggest Bank Failure in American History.* New York: Simon & Schuster.

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

Hetzel, Robert L. 2012. *The Great Recession: Market Failure or Policy Failure?* New York: Cambridge University Press.

Kuprianov, Anatoli. 1995. "Derivatives Debacles: Case Studies of Large Losses in Derivatives Markets." Federal Reserve Bank of Richmond *Economic Quarterly* 81 (Fall): 1–39.

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

Marshall, Elizabeth, Sabrina R. Pellerin, and John R. Walter. 2013. "2013 Estimates of the Safety Net" (Using Data as of Dec. 31, 2011). Available at www.richmondfed.org/publications/research/special_reports/safety_net/pdf/safety_net_methodology_sources.pdf.

McLean, Bethany, and Joe Nocera. 2010. *All the Devils are Here: The Hidden History of the Financial Crisis.* New York: Penguin Group.

Merrill Lynch & Co., Inc. 1998. "1998 Annual Report." Available at www.ml.com/annualmeetingmaterials/annrep98/index.htm.

Merrill Lynch & Co., Inc. 2006. "2006 Annual Report." Available at www.ml.com/annualmeetingmaterials/2006/ar/.

Merrill Lynch & Co., Inc. 2007. "2007 Annual Report." Available at www.ml.com/annualmeetingmaterials/2007/ar/.

Merrill Lynch & Co., Inc. 2007. "Form 10-Q for the Period Ending June 29, 2007." Available at http://phx.corporate-ir.net/phoenix.zhtml?c=93516&p=irol-sec.

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

Rothacker, Rick. 2010. *Banktown: The Rise and Struggles of Charlotte's Big Banks.* Winston-Salem, N.C.: John F. Blair.

Sprague, Irvine H. 1986. *Bailout: An Insider's Account of Bank Failures and Rescues.* New York: Basic Books, Inc.

Stern, Gary H., and Ron J. Feldman. 2009. *Too Big to Fail: The Hazards of Bank Bailouts.* Washington, D.C.: Brookings Institution Press.

UBS. 2008. *Shareholder Report on UBS's Write-Downs.* Zurich: UBS (April 18).

White, Lawrence J. 1991. *The S&L Debacle.* New York: Oxford University Press.