Financial Factors in the Great Depression

Charles W. Calomiris

Beginning with Irving Fisher (1933) and John Maynard Keynes (1931 [1963]), macroeconomists have argued that financial markets were important sources and propagators of decline during the Great Depression. Turning points during the Depression often coincided with or were preceded by dramatic events in financial markets: stock market collapse, waves of bankruptcy and bank failure, and contractions in the money stock. But the mechanism through which financial factors contributed to the Depression has been a source of controversy, as has been the relative importance of financial factors in explaining the origins and persistence of the Depression.

This essay reviews the literature on the role of financial factors in the Depression, and draws some lessons that have more general relevance for the study of the Depression and for macroeconomics. I argue that much of the recent progress that has been made in understanding some of the most important and puzzling aspects of financial-real links in the Depression followed a paradigm shift in economics. A central, neglected theoretical piece of the story for financial factors was the allocative effects of imperfections in capital markets, which can imply links between disruptions in financial markets and subsequent economic activity. Also, the increasing emphasis on learning and “path-dependence” in economics has helped to explain why financial shocks during the 1930s were so severe and why policy-makers failed to prevent the Depression.

Charles W. Calomiris is Associate Professor of Finance, University of Illinois, Urbana-Champaign, Illinois, and Faculty Research Fellow, National Bureau of Economic Research, Cambridge, Massachusetts.
The Monetarist Revolution and the Great Depression

In their monumental *Monetary History of the United States* (1963), Milton Friedman and Anna Schwartz provided a simple and potentially powerful explanation of the origins of the Great Depression that depended on exogenous changes in the money supply. Just as importantly, they clearly defined crucial elements of the sequence of events from 1929 to 1940 that any theory of the Depression would have to explain, especially the co-movements of nominal GNP and the money stock, the movements of prices, and changes in the relative size of various components of the money stock.

Friedman and Schwartz were less interested in explaining the beginnings of the recession of 1929 and the October stock market crash than in the question of how an initial downturn in 1929 became transformed into the Great Depression. They argued that waves of banking crises, beginning in October 1930 and ending in March 1933, substantially reduced the money multiplier and the money stock. The failure of the Federal Reserve to offset this decline with open market operations and loans to banks through the discount window led to a drastic contraction in economic activity. They argued this policy failure resulted from a change in leadership within the Fed (notably, the departure of Benjamin Strong). Monetary ease and recovery from 1933–1936 was followed in 1937 by contractionary monetary policy and economic decline, which Friedman and Schwartz traced to the doubling of the required reserve ratio in an ill-conceived attempt to reduce excess reserves in the banking system.¹

Since its publication, Friedman and Schwartz’s *Monetary History* has defined much of the research agenda for the study of connections between financial markets and real activity during the Depression. Subsequent research continues to address five broad categories of questions raised directly or indirectly by Friedman and Schwartz’s work:

1) To what extent are the reductions in the money supply from 1930 to 1933, and the waves of bank failure that Friedman and Schwartz focused on to explain them, properly viewed as exogenous to the decline in income, and to what extent were they merely symptoms of a decline that had separate origins?

2) In light of the near-zero nominal short-term interest rates of the 1930s, was it possible to argue that the demand for money was stable, and that a decline in money supply would lead to a fall in nominal income? Or was there an elastic demand for money at low interest rates (that is, a Keynesian liquidity trap)?

¹The attempt to reduce excess reserves resulted from the mistaken belief that excess reserves were an unnecessary surplus and a potential threat to monetary control. Friedman and Schwartz (1963) argued that high excess reserves reflected increased liquidity preference by banks in the face of the crisis in the financial system.
3) Were Fed actions and the failure of policy during the "Great Contraction" of the money stock the result of new policy and new leadership, as Friedman and Schwartz contended, or did they represent the application of the same old formulas to new circumstances?

4) Could a monetarist explanation, or any explanation relying on nominal price and wage rigidity, account for the persistent stagnation of the economy during the 1930s?

5) Could Federal Reserve open market operations alone, unaccompanied by changes in monetary and regulatory regimes (like the departure from the gold standard, direct government intervention to assist banks, or the suspension of convertibility requirements for deposits), have reversed economic decline at any time during 1930–1933, as Friedman and Schwartz claimed?

For two decades after the publication of Monetary History, the literature on the Great Depression focused on the first three of these questions. Roughly speaking, economists agreed that the sticky-price, IS-LM paradigm was the proper framework within which to capture the links between financial and real markets, although they disagreed on some details, like which interest rate to focus on as a measure of monetary stringency in 1930–1933, and on how to think of the adjustment process toward market clearing in the goods, bond, and money markets. A few dissidents saw the neoclassical synthesis as inherently incapable of capturing real-financial links (Gurley and Shaw, 1960; Goldsmith, 1969; McKinnon, 1973; and Minsky, 1975 are noteworthy), but only Kindleberger (1973) focused on the Depression. His insistence on complex financial linkages and feedback across countries, without supplying formal modeling or measurement of these mechanisms, was welcomed with the enthusiasm accorded Banquo at Macbeth’s feast.

Because Friedman and Schwartz, and their supporters and critics, initially framed their debate more or less within the standard IS-LM model, financial shocks were viewed through the narrow windows of changes in the money stock, or stock-market price effects on wealth, and hence, consumption demand. The consensus achieved by this literature initially was limited.

Progress was made on the narrow question of whether monetary shocks could have been an important source of disturbance during the 1930s. Various researchers found that money demand was stable during the 1930s (that there was no liquidity trap), and hence, that money-supply shocks could have had important effects on output (Meltzer, 1963; Gandolfi, 1974; Gandolfi and Lothian, 1977).

Nevertheless, others questioned the exogeneity of money-supply changes or their importance during the banking crises, and noted that the real stock of money had not contracted during the early stages of the Depression (as shown in Figure 1) as should have occurred, in the context of an IS-LM model, if money supply had been the dominant source of disturbance (Temin, 1976; Gordon and Wilcox, 1981). Critics advocated additional "autonomous-
Figure 1
Money, Prizes, Production

Source: Bank deposits are from Friedman and Schwartz (1970, pp. 507–13, columns 3 + 6), the wholesale price index and industrial production index are from U.S. Department of Commerce (1949, pp. 344, 310)

expenditure” shocks to explain the origins of the Depression (Hickman, 1973; Temin, 1976; Gordon and Wilcox, 1981). Meltzer (1976) argued that international monetary forces, driven by misalignment of prices across countries, may have caused early price and output reductions through the price-specie-flow mechanism.²

These early debates about the sources of disturbances continue and much remains unsettled (Bordo, 1986). Conclusions about the relative importance of monetary and autonomous-expenditure shocks have turned out to be quite sensitive to empirical methodology and different researchers’ interpretations of observed time-series relationships. Large, autonomous consumption reductions in 1929–1930, posited by Temin (1976), have been confirmed by Hall (1986) and Romer (1990), but questioned by Gordon and Wilcox (1981), Gordon and Veitch (1986), and others. Gordon and Wilcox (1981) find that the association between lagged money and current income is weak for the 1930s, while the association between contemporaneous movements is stronger, which they argue is more consistent with endogeneity of money. Monetarists respond that the relationship between money and GNP is subject to lags of variable and

²According to Hume’s price-specie-flow mechanism, international price disequilibrium brings forth endogenous changes in international flows of goods, and offsetting flows of specie, which realign price levels across countries.
uncertain length, and cite evidence that banking crises in the 1930s are associated with changes in the money multiplier (for example Anderson and Burkiewitz, 1980; Boughton and Wicker, 1979; Schwartz, 1981; Trescott, 1984). As Wicker (1989) points out, however, the association between bank failures and money-multiplier changes need not imply exogenous change in the money supply, since both may have followed income and interest rate changes. Indeed, the most likely exogenous source of change in the money stock on a priori grounds, the first banking crisis of 1930, may have been primarily of local importance and seems to have had little effect on national economic activity (Friedman and Schwartz, 1963, p. 313; Wicker, 1980, 1982), although it did mark a change in the risk premium for low-grade corporate securities (Hamilton, 1987). In a similar vein, White (1984) argues that the first banking crisis of 1930 was not an exceptional event, or a turning point for the banking system, but rather represented a continuation of patterns of bank failure during recession experienced in earlier years.

Recently, some convincing evidence has emerged of exogenous disturbances in the market for money balances during the Depression. Ironically, the clearest evidence produced on the importance of exogenous changes in the money supply pertains to the pre-October 1929 period (which had always been viewed by monetarists and non-monetarists alike as a period of tight monetary policy), and to subsequent influences on the money supply more moderate than the sharp contractions of the money stock during the banking crises emphasized by Friedman and Schwartz. Field (1984a, 1984b) showed that securities market trading increased the demand for money in the late 1920s, and that this increase in demand was not offset by expansion in supply. Indeed, the expansion in demand worked in concert with the contraction in money supply in 1929 to increase interest rates and reduce prices and economic activity. Wheelock (1990) found that a subsequent downward shift in banks' demand for borrowed reserves caused a reduction in the money multiplier which the Fed did not offset with open market operations. This produced a persistent reduction in the money supply during the Depression. While of interest, these studies provided little direct support for the central and truly novel point of Friedman and Schwartz's thesis—that waves of severe exogenous monetary contraction beginning in late 1930 converted the relatively normal recession of 1929–1930 into the Great Depression collapse of 1930–1933.

The criticisms of Friedman and Schwartz deprived monetary shocks of their status as primary, indisputable forces in the Depression. The Friedman-Schwartz view, while coherent as an explanation of the fall in income from 1930 to 1933, lacked empirical evidence that could not be explained by other reasonable interpretations of the data. Even staunch advocates of monetarism (like Meltzer, 1981) retreated to compromise positions in light of the new evidence, and focused instead on the counterfactual point—that stable money demand implied that the Fed could have prevented the Great Depression, if policy had been wiser.
Even this weak form of the Friedman-Schwartz argument—that the Fed should have done a better job conducting monetary policy, if only in reaction to other exogenous events—was undergoing challenge by economic historians. Friedman and Schwartz had argued that, had he lived, Benjamin Strong would have done a much better job managing policy than his successors. Judging by the standard of past Fed policy, they argued that the policies of 1929–1933 represented a movement backward in competence. This was an important argument for Friedman and Schwartz. If Federal Reserve policy had been ineffectual or counterproductive throughout the interwar period, then one could not reasonably argue that effective monetary policy was part of the available “technology” at the time of the Depression. This issue is central to the question of whether the Depression was avoidable at the time.

Elmus Wicker (1965) was the first to raise objections to Friedman and Schwartz’s view of changes in Federal Reserve targeting in the 1930s, and his views were buttressed by Brunner and Meltzer (1968). In essence, these and other critics argued that the Federal Reserve did not change policy regime in the 1930s, that policy was often unwise or ineffective, and that the Fed’s behavior prior to 1933 was constrained by poor targets and indicators (stock prices, borrowed reserves, gold flows, and interest rates), poor understanding of the economy, and by an adherence to the gold standard and a consequent emphasis on international as well as domestic objectives. As part of maintaining the gold standard, central banks must respond eventually to persistent outflows of gold with contractionary open market operations, to drive up interest rates, attract gold, and preserve gold reserves. Thus the Fed’s pursuit of domestic objectives was limited by its commitment to maintain a credible long-run link to gold, and by its view of what policy responses that entailed. Furthermore, open market operations had little overall effect on the supply of high-powered money because they often were offset by changes in member bank borrowings (Toma, 1989). Wheelock (1989a, 1989b, 1992) provided supporting descriptive and econometric evidence for the stability of the Federal Reserve’s reaction function over the interwar period, which caused the Fed to misread credit conditions in early 1931, and to fail to expand the money supply in late 1931, even after it became aware of tight credit-market conditions. Consistent with its long-standing policies, in early 1931, the Fed interpreted high excess reserves and low interest rates—which were the result of a massive worldwide flight to liquidity by individuals and banks—as signs of easy money, which warranted higher interest rates to preserve external balance. After Britain left gold in September 1931, outflows of reserves from the United States prompted tightening of monetary policy to preserve external balance.

This reaction function had been derived from previous experience and by the prevailing doctrines of central bank policy under the gold standard (Tomin, 1989). This policy may have been appropriate in some circumstances, but it increased the fragility of the financial system and contributed to the decline of money, credit, economic activity and prices in 1931. One can lament poor
policy by the Fed in the 1930s, and it is true that some criticized Fed policies at the time (thus wise advice was available, in principle), but one cannot expect the Fed to have learned the lessons of the Great Depression before it happened.

A Change in Paradigm, New Questions, and Old Answers

For two decades after *Monetary History*, the literature on the Great Depression argued cause and effect essentially within the confines of the neoclassical synthesis that reigned in macroeconomics in the 1960s and 1970s. In this context, financial factors are identified primarily with money-supply shocks and stock market influences, which in turn affect investment and consumption demand through interest elasticities, wealth effects, and changing perceptions of uncertainty (Temen, 1976; Gordon and Vetsch, 1986; Romer, 1990).

But a transformation in thinking about the role of financial markets in the economy was under way. Economists began to formulate theoretical arguments of why conditions in financial markets might not be accurately captured by the aggregate value of capital in the stock market, the supply of money, and "the" real or nominal interest rate. Theoretical models of credit allocation under asymmetric information imply that access to external finance may be inhibited because of information costs faced by sources of outside funding. Under these circumstances, "insiders"—firm managers and financial intermediaries with an ongoing relationship with the firm—can supply funds at lower cost than "outsiders"—relatively uninformed stockholders and bondholders. An important implication of this literature is that changes in the allocation of wealth in the economy can increase the cost of outside finance if they reduce the available supply of "insider" funding. For example, decreases in the wealth of insider shareholders, or reductions in bank net worth that inhibit bank lending capacity, will increase firms' reliance on outside funds and drive up the cost of those funds. Furthermore, the demands for assets and the pricing of assets will reflect the extent to which assets are "liquid"—that is, the extent to which their value is a matter of common knowledge. Early contributions to this literature included Akerlof (1970), Jaffee and Russell (1976), Leland and Pyle (1977), Stiglitz and Weiss (1981), and Myers and Majluf (1984).

Mishkin (1978) was the first to apply the new literature on imperfect capital markets to the Great Depression. Mishkin (1976) presents a model of consumer "distress" to analyze the role of debt deflation in reducing consumer durables demand. He argues that consumers valued "liquidity" (that is, holding wealth in assets that do not suffer distress-sale discounts due to asymmetric information about their true value). Exogenous shocks to consumer liquidity will lead consumers to reduce their demand for illiquid consumer durables as they try to rebuild their stock of liquid assets. This framework served as the basis for Mishkin's (1978) study of the effects of changes in the household balance sheet and consumer expenditures during the Depression. According to
Mishkin (1976), the changing distribution of wealth, not just aggregate wealth, should matter for aggregate consumption. Mishkin (1978) argued that in addition to the depressive effect of aggregate wealth reduction on consumption in the 1930s, the debt deflation (a reallocation of wealth away from indebted consumers) reduced aggregate consumption demand.

Mishkin's research, with its emphasis on the depressive effects of excess leverage and the allocative consequences of wealth redistribution in the presence of capital market imperfections, marked an important change in the direction of the literature on financial factors in the Depression. However, his contribution still remained within the confines of the neoclassical synthesis, as part of the explanation for the early autonomous contraction in consumption demand. Bernanke's (1983) study of the consequences of financial disruption during the Depression took Mishkin's arguments a step further. Bernanke argued that in the presence of capital market imperfections, the destruction of intermediaries and the reduction in borrowers' net worth—both the results of debt deflation—reduced investment in the 1930s by increasing the marginal cost of funding. Reductions in firm net worth increase credit costs for firms because as debt deflation erodes the equity stake of firm "insiders," the ratio of external to internal claims on the firm rises. Under asymmetric information this increases the marginal cost of external finance. Debt deflation also erodes the net worth of banks, causing some banks to fail, and others to tighten their credit standards to avoid runs by depositors. Thus access to "inside" debt from relatively well-informed banks is also curtailed. The financial devastation of 1929 to 1933 had always been given prominence in accounts of the Depression. Bernanke's (1983) contribution was to combine theory and empirical evidence to argue that financial collapse was more than a symptom of economic decline; financial collapse deepened the Depression by hampering the efficient allocation of capital.

In retrospect, given the dramatic changes that occurred in financial markets from 1929 to 1933, it may seem surprising that it took so long to develop such an argument. During this period the ratio of high-to-low rated bond debt fell from 2.4 to 0.3 (Hickman, 1960, p. 21). Defaults on bonds from 1930 to 1939 were nearly triple the number that had occurred from 1920 to 1929 (Hickman, p. 249), and the market value of defaulted issues (for 1930-1943) at their default dates averaged 34 percent of par, compared to 61 percent for the period 1920-1929 (Hickman, p. 560). As shown in Figure 2, the quality spread in bond market yields jumped dramatically during the Depression, and real liabilities of failed businesses tripled. Nominal liabilities of failed businesses rose from a monthly average of $40 million for January 1928 through December 1929 to a monthly average of $63 million for January 1930 through June 1933 (U.S. Department of Commerce, 1949, p. 349).

Moreover, bank failures rose to historically unprecedented heights, with historically unprecedented costs to depositors. The banking collapse of the 1930s differed in kind and degree from earlier banking crises. During the
national banking era, nationwide bank suspensions of convertibility by banks occurred at, or just following, cyclical peaks. These suspensions were short-lived and produced few bank failures (Calomiris and Gorton, 1991). In the 1930s, widespread suspensions came late in the cycle, in 1931 and 1933, after banks had suffered devastating losses due to borrower bankruptcies and deflation. Destabilizing deflation was fueled by persistent withdrawals of deposits and contraction in the money supply, which could have been prevented by an early nationwide suspension of convertibility. Unlike those of earlier periods, bank suspensions during the Depression were not brief and resulted in unprecedented numbers of failed banks. The period from 1921 to 1929, which itself saw an unusual rate of loss for banks due to the agricultural depression of the 1920s, paled by comparison to the much shorter period from 1930 through 1933. For 1921–1929, the deposits of failed banks totaled $1.6 billion, with estimated losses to depositors of $565 million. For 1930–1933, the deposits of failed banks totaled $6.8 billion, with estimated losses to depositors of $1.3 billion (Board of Governors, 1943, p. 283). Surviving banks substantially curtailed their lending, with loan-to-deposit ratios falling from 0.85 in 1929 to a low of 0.58 in January 1933.

5 Nearly half of the liabilities of suspended banks for the period 1930–1933 is attributable to suspensions that coincided with the bank holiday of 1933 (Bernanke, 1983, p. 262).
Bernanke's research was informed by new models of capital market failure, but these were not the main motivation for his approach to modeling connections between financial collapse and real decline in the 1930s. The search for a new paradigm of financial-real interaction followed from an incompleteness in the earlier literature on the Depression, which Bernanke focused on in his introduction (p. 257):

One problem is that there is no theory of monetary effects [per se] on the real economy that can explain protracted nonneutrality. Another is that the reductions of the money supply in this period seems quantitatively insufficient to explain the subsequent falls in output.

According to Bernanke, the decline in the efficiency of the economy's financial allocation mechanism induced by the reduction of banks' lending capabilities and the collapse of producers' and consumers' net worth should be thought of as long-lived shocks to financial technology, and therefore, can explain the persistent decline in output through a rise in the "cost of credit intermediation." Indeed, if shocks to credit costs mainly constrained the growth of newer, technologically innovative, "information-intensive" firms with relatively less access to credit facilities in the 1930s, the long-term consequences for economic activity might have been especially pronounced (Hunter, 1982; Calomiris and Hubbard, 1991).

Another weakness in the IS-LM approach to modeling real financial links during the Depression was its dependence on price stickiness. Explanations of economic decline during the Depression that rely on reductions in real money balances, autonomous changes in expenditure, or a price-specie-flow mechanism all assume price rigidity or price disequilibrium. But as Figure 1 shows, wholesale prices and bank deposits show close contemporaneous co-movements even at high frequencies, which argues against the assumption of protracted price adjustment, at least for wholesale prices.4

Bernanke cited (and likely was motivated by) the writings of several 1930s chroniclers of credit market conditions and economists who emphasized persistent disruption to financial markets as one of the main continuing problems of the Depression after 1933, and who viewed deflation as a destabilizing influence.5 The economists on this list include Irving Fisher (1933), whose classic

---

4 It is also interesting to recall that Gordon and Wilcox (1981) found the strongest association between money and economic activity was essentially contemporaneous. Calomiris and Hubbard (1989) found similar results for financial-real association during the pre-World War I period and argued that strong contemporaneous association made more sense in the context of a credit squeeze (in which flows of goods may be abruptly discontinued) than in a standard monetary disequilibrium story, which should involve a protracted process of portfolio and price adjustment.

5 Calomiris and Hubbard (1989) argue that chroniclers and economists before Friedman and Schwartz generally used the phrase "money market" to mean the market for short-term credit, and that pre-Federal Reserve real-financial links were properly seen by contemporaries as the result of shocks to credit supply (partly involving unanticipated deflation).
statement of the debt-deflation cycle mirrors many of Bernanke's arguments. Perhaps the most prominent advocate of this position in the 1930s was Keynes (1931 [1963], pp. 175–76) who wrote:

...there is scarcely any class of property, except real estate, however useful and important to the welfare of the community, the current money value of which has not suffered an enormous decline. This has happened in a community which is so organised that a veil of money is, as I have said, interposed over a wide field between the actual asset and the wealth owner. The ostensible proprietor of the actual asset has financed it by borrowing money from the actual owner of wealth. Furthermore, it is largely through the banking system that all this has been arranged. That is to say, the banks have, for a consideration, interposed their guarantee. They stand between the real borrower and the real lender. They have given their guarantee to the real lender; and this guarantee is only good if the money value of the asset belonging to the real borrower is worth the money which has been advanced on it. It is for this reason that a decline in the money values so severe as that which we are now experiencing threatens the solidity of the whole financial structure.

In his empirical work, Bernanke showed that both shocks to firms' and banks' net worth were significant (statistically and economically) for explaining the fall in output during the 1930s, even after taking account of monetary shocks. Indicators of declining net worth of banks and firms—including deflation, corporate failures, bank failures, and the bond risk spread—all were important as predictors of economic decline over and above monetary shocks.

The new focus on deflation and financial disruption also had implications for the way the Depression was transmitted across countries. Indeed, Kindleberger's (1973, pp. 144–45) analysis of international transmission had argued that stock market decline and deflationary shocks, which precipitated a liquidity squeeze, a contraction in bank lending, and international financial collapse beginning in 1930, turned the recession of 1929–1930 into the Great Depression:

New lending stopped because of falling prices, and prices kept falling because of no new lending. As the less-developed countries lost access to loans and spent their gold and foreign-exchange reserves, they were forced to sell old quantities of primary products for what the market would bring. Deflation spiraled.

In cross-country comparisons, Kindleberger (1973, pp. 232 ff) also emphasized that countries remaining on gold after 1931 suffered from deepening depression, while those that abandoned gold began the process of recovery. By remaining on gold, countries tied their price levels to a declining world trend
as the demand for gold kept rising. Countries leaving gold were free to pursue independent monetary policy and bring an end to deflation. International comparisons linking maintenance of the gold standard, deflation, and continuing decline have been confirmed in more formal studies (Eichengreen and Sachs, 1983, 1986; Temin, 1989; Bernanke and James, 1991; Eichengreen, 1992).\footnote{Despite these elements of agreement, there remain important differences among these authors regarding the desirability of a worldwide departure from gold. Kindleberger (1973, pp. 294-7) argues that the failure to support the gold standard early on through coordination of central bank policies (an international lender of last resort) caused worldwide deflation, while Eichengreen (1992, pp. 301-302) takes exception to this view and claims that absent fundamental long-run changes in policy, maintenance of the gold standard was untenable. From Eichengreen's perspective, the departure from gold was a necessary step to avoid the Depression, while Kindleberger views the collective abandonment of gold as a mistake.}

The "new view" of Bernanke and others was not a rejection of Friedman and Schwartz's argument that monetary shocks were important. Its main contribution was to show that monetary shocks, and other disturbances during the early phase of the Depression, had long-run effects largely because they affected the institutional structure of credit markets and the balance sheets of borrowers. Indeed, Hamilton (1987) argued that the appropriate model of the origins and persistence of the Great Depression combines monetary shocks to explain the origins of the recession of 1929, and other shocks in 1930, with consequent unanticipated deflation from 1930 to 1933, which operated on the economy through Bernanke's transmission mechanism. Based on his analysis of futures market prices, he argued that deflation was unanticipated, and that increases in the bond risk spread coincided with the onset of financial disruption in 1930. Hamilton (1987, 1992) does not resolve the question of whether autonomous consumption or money-supply shocks were more important in 1930; he argues that substantial deflation followed these shocks, and that deflation had persisting influences.

In retrospect, despite their focus on the money stock, Friedman and Schwartz drew attention to evidence favorable to the "new view." Friedman and Schwartz (1963, pp. 312-15) and Schwartz (1981, pp. 31-38) saw high yield spreads on bonds as indicating a general liquidity crisis which they associated with a fall in the money multiplier in late 1930. This emphasis on broader definitions of "liquidity" than the available money stock is consistent with the direction pursued by Bernanke, Hamilton, and Mishkin in their research. In particular, Mishkin (1991a, 1991b) argued that reductions in stock prices and increases in quality spreads in bond markets are best viewed as the result of changes in "lemons" discounts on securities prices under asymmetric information, which signal financial-market disruption and reductions in internal funds available to firms.\footnote{Formal models with these implications abound (Myers and Majluf, 1984; Greenwald and Stiglitz, 1988; Bernanke and Gertler, 1989, 1990; Calomiris and Hubbard, 1990; Brock and LeBaron, 1990; Gertler and Hubbard, 1991).} Financial market disruption has a larger impact
on certain firms, and thus reduces their creditworthiness and the value of their securities in the market. Mishkin (1991a) analyzed yield spreads for bonds and stock price movements during historical financial panics and confirmed a "flight to quality" (that is, an increased penalty for firms whose prospects suffer an asymmetric-information discount). He applied this analysis to bond and stock markets in 1930, and argued for a similar "flight to quality" at the beginning of the recession of 1937. Mishkin’s analysis of securities markets provides a theoretical explanation for time variation in the risk premium of stocks and bonds by linking these changes to exogenous disruptions in financial markets throughout U.S. history. 8

Implications of the “New View”

The non-monetary propagation hypothesis has at least three interesting implications that distinguish it from the earlier monetarist position.

First, the financial-propagation view of the Depression implies that a money-supply shock of a given magnitude will have a larger effect if it occurs at a time of high leverage, or in an economy with a poorly diversified, geographically fragmented banking system like that of the United States. From this perspective, the “new-age” optimism of the 1920s (Domínguez, Fair, and Shapiro, 1988; Romer, 1990; Nelson, 1991) worsened the magnitude of the financial and real reactions of the 1930s in the face of deflationary shocks. That optimism had been reflected in stock price rises (White, 1990; Rappaport and White, 1991; De Long and Shleifer, 1991) and debt accumulation, notably in the new consumer debt market (Mishkin, 1978). The effect of the pre-existing debt burden on the balance sheets of consumers and producers may have been somewhat muted in the early years of the Depression by an expectation of an early recovery. However, the unprecedented debt burden, which continued to rise during the deflation of the early 1930s (as shown in Figure 3), reduced the creditworthiness of many borrowers through drastic redistributions of wealth. The extent to which the Depression’s severity was a consequence of the financial boom that preceded it remains an interesting question for future research.

Second, once the character of financial market influences is broadened to include non-monetary channels, earlier monetarist arguments that restoring the money supply to earlier levels through open market operations in 1931-1933 could have reversed the course of the Depression and prevented bank failure, borrower insolvency, and economic decline must be qualified. Because the financial system is path-dependent, disturbances to the allocation of wealth and the viability of financial intermediaries caused by open market

8Of course, one could argue that time-varying risk premia at times of financial crisis reflect other influences. Disentangling the relative contribution of asymmetric information and other cyclical influences on risk premia remains an important topic for future research.
Figure 3
Real Debt and Stock Prices

Source: The GNP deflator is from Romer (1989, p. 23) for 1920–1929, and Balke and Gordon (1986, p. 782) for 1930–1935, private long-term debt is from Kuzin (1986, p. 36, Table 10, column 1); the stock price index is the average of quarterly data from Balke and Gordon (1986, pp. 803–4)

operations cannot in general be reversed by open market operations that restore the money supply. Borrowers or bankers who have already suffered large losses due to increased debt burdens and costs of financial distress will not regain lost wealth as the result of subsequent open market operations. Their balance sheet positions have changed. Moreover, the effects of policy-induced deflation and inflation are asymmetric; that is, too much leverage is penalized by the capital market by more than too little leverage is rewarded, and costs of financial distress incurred in states of low net worth have no counterpart in states of high net worth. Thus, according to the new view, even if Friedman and Schwartz and their supporters were entirely correct about the importance of monetary shocks in precipitating the Depression, it does not follow that open market operations to restore the money supply would have had offsetting effects in promoting recovery from the Depression.

Third, different methods of increasing the money supply—say, expansionary open market operations vs. reductions in the discount rate—might have had very different consequences for recovery (contrary to Friedman and Schwartz, 1986, p. 201). The discount window could have been used to provide

6Furthermore, if expansionary open market operations to reverse a previous decline are anticipated, they would have no effect on the allocation of wealth, and thus would not reverse the effects of previous unanticipated deflation.
focused assistance to the banking system and decrease the relative cost of bank credit, while the benefits of expansionary open market operations would have been confined to increases in the aggregate supply of money, increased prices, and reduced interest rates on riskless short-term securities. This argument certainly does not mean that subsidies to firms, deposit insurance, or bank bailouts are always a desirable alternative form of government intervention; however, one of the implications of the new financial view is that open market operations may be a blunt and insufficient instrument for reversing the effects of bad previous policy compared to other policies. This reasoning underlay the arguments of some financial analysts who advocated other policies, including direct government assistance to banks and firms through the Reconstruction Finance Corporation, to stimulate recovery from the Depression (Clark, 1933; Harris, 1933, p. 719).

Was Deflation Unanticipated?

Several recent papers have challenged or supported the new synthesis of early monetary (and other) shocks and long-run financial propagators for explaining the origins and persistence of the Depression.

A central element of the new financial view is that deflationary shocks from 1929 to 1933 were largely unanticipated (otherwise, they would not have produced financial distress), and this has been the topic of several papers. In support of Hamilton's (1987) results, several papers argue that relatively sanguine expectations and forecasts of economic activity and prices persisted into 1929 and 1930 (Dominguez, Fair, and Shapiro, 1988; Romer, 1990; Nelson, 1991). Hamilton's (1992) subsequent work has reiterated his earlier findings using new methods for extracting price forecasts from futures prices.

While there is some continuing disagreement between these authors and Cecchetti (1992) over precisely how much of the deflation was anticipated at short time horizons, all parties agree that there was substantial unanticipated deflation even at quarterly frequencies. Furthermore, Evans and Wachtel (1991) argue that over longer-term frequencies most of the deflation was unanticipated. They find that agents systematically overestimated the probability of a return to a zero-inflation regime rather than continuing deflation. Given that debt contracts often were written with durations greater than several months, the rise in the real value of long-term debt that occurred must have been unanticipated. Kuhin (1936, p. 36) estimates total long-term debt as roughly $84 billion in 1929 (at a time when national income was roughly $87 billion, and short-term debt was roughly $150 billion, as estimated respectively

19Anticipated deflation still could have had a depressive effect on the economy, as argued by Temin (1988); but it would not have produced financial distress, since agents anticipating deflation would reduce interest costs to offset the capital loss from debt deflation.
by Kuznets, 1941, v. 1, p. 147, and Clark, 1933, p. 301). Hickman (1960, p. 60) reports that 98 percent of outstanding bonds in 1928 had maturities of greater than one year, and 78 percent had maturities of greater than 5 years. Mortgages, which comprised half of both farm and nonfarm household indebtedness in 1929 (Goldsmith, 1962, v. 3, pp. 67, 75), typically had durations of three to five years (Snowden and Bu-Saba, 1992).

New Challenges and Interpretations

The new view's emphasis on deflation prompted several cross-country comparisons of the role and transmission of deflation. In notable recent work, Haubrich's (1990) study of financial-real interactions during the Depression applies empirical methods similar to Bernanke (1983) to Canada. He finds that measures of financial distress have no economic or statistical significance for predicting economic activity in Canada. Haubrich interprets this as evidence that without bank failures (which were absent in Canada's increasingly concentrated nationwide branch-banking system), financial distress has little macroeconomic consequence. However, there are at least four reasons to doubt the general proposition that macroeconomic financial distress depends on widespread bank failures, and the application of this proposition to the 1930s.

First, Canada and the United States were not equally vulnerable to financial disturbances at the time of the Depression. According to theoretical models of the allocative effects of wealth redistribution, the effects on economic activity of deflation-induced reductions in net worth are nonlinear and depend on the initial balance sheet position of the firm, and its initial composition of inside and outside funding (Bernanke and Gertler, 1989, 1990; Calomiris and Hubbard, 1990). For example, a borrower with a low debt-to-asset ratio or a large preexisting amount of inside equity funding may remain highly creditworthy even in the face of severe deflationary shocks. While the United States and Canada suffered similar deflations from 1929 to 1933, Canadian real debt burdens started lower and never reached levels comparable to the United States. In the United States from 1929 to 1933, debt service relative to GNP increased from 9 percent to 19.8 percent, while in Canada it rose from 3.9 percent to 6.4 percent (Haubrich, 1990, p. 242). As noted earlier, the run-up in debt burdens during the 1920s in the United States set the stage for the deflation-induced increase in financial distress; in Canada, the run-up in the 1920s was less pronounced, and thus one would expect the impact of financial shocks to have been weaker.

Second, in Canada the money stock may have been a better indicator of the outstanding volume of short-term credit than in the United States where non-bank forms of credit (like commercial paper) were much more important. Thus in comparable regressions that include the money stock, one would expect changes in the Canadian money stock to capture changes in the cost of credit better than money stock changes in the United States, leaving more of a
"credit-cost" residual to be explained by non-monetary factors in the United States. Thus financial factors may have been important in Canada, but less likely to show up as significant in regressions that also include the money stock.

Third, Canada's relative reliance on banks as sources of short-term credit, and the concentration of the banking industry, may have reduced the costs of managing financial distress in Canada relative to the United States. The concentration of lending and renegotiation authority may have reduced the impact of increased debt burden on economic activity. Recent studies have found that both the concentration of lending and the reliance on banks for loans mitigate declines in firms' securities prices and funding sources during financial distress (Hoshi, Kashyap, and Scharfstein, 1990a; Gilson, John, and Lang, 1990; Brown, James, and Mooradian, 1991).

Fourth, Haubrich's interpretation of his findings implies that exogenous variation in "inside equity" has smaller allocative consequences than similar variation in the availability of "inside debt" (bank loans). It is difficult to justify this distinction between inside debt and inside equity as a theoretical proposition. The central point of the asymmetric-information approach to corporate finance is that outside funds, whether debt or equity, entail greater costs than funds supplied by relatively informed stockholders/managers and their bankers. At times when cash flow is scarce, or when insiders' stakes are reduced, the cost of funds rises. Recent empirical findings in the literature on investment and corporate finance support this approach to firms' costs of funds (Fazzari, Hubbard, and Petersen, 1988; Hoshi, Kashyap, and Scharfstein, 1990b; Mackie-Mason, 1990). Calomiris and Hubbard's (1991) study of American manufacturing firms in the mid-1930s shows that accumulated retained earnings were a substantial constraint on corporate investment for roughly one quarter of firms. It follows that at least these firms would have substantially reduced their investment in response to a deflation-induced reduction in inside equity.

For these reasons I do not think Haubrich's findings support his general conclusion (and his suggested interpretation of Bernanke, 1983) that bank failures are a necessary precondition for the transmission of financial distress. Rather, Haubrich's study suggests (subject to the caveat of my second qualification of his results) how different countries' financial institutions and initial conditions affect their relative vulnerability to financial disturbances.

Temin (1989) challenges the importance of unanticipated deflation and Bernanke's financial transmission mechanism for the United States. Temin argues that the anticipated component of the deflation (post-1930) must have been more important, particularly downward rigidity of nominal interest rates (which keeps the cost of borrowing high), and the Mundell-Tobin portfolio reallocation effect.¹¹ Temin rejects the Bernanke-Hamilton view because, he

¹¹The Mundell-Tobin effect relies on portfolio allocation toward money when the inflation rate is low (Mundell, 1963; Tobin, 1966). For example, in Tobin's dynamic framework, expected deflation reduces the attractiveness of holding real capital and thus reduces equilibrium economic activity.
argues, one of its major predictions fails to hold: firms with relatively high costs of external finance (small firms) should have suffered the most from the increase in the cost of credit intermediation. Temin constructs a test of this proposition and rejects it. He divides the economy into industries and asks whether industries with low concentration ratios suffered unusually severe contractions relative to other industries compared to other cyclical downturns. Finding no such pattern, he rejects the importance of increases in the cost of credit as the propagator of deflation.

I would raise three objections to Temin’s test. First, differences across firms in costs of finance may not show up in industry-level aggregation. Second, concentration ratios at the industry level may be a very poor indicator of cross-industry variation in external finance costs. Firm size, which itself is only indirectly related to Temin’s measure, is an imperfect indicator of finance costs, and Calomiris and Hubbard (1991) show that it is a very noisy indicator for the mid-1930s. Third, cross-sectional differences in industry performance may be hard to observe during a massive disturbance like the Depression that substantially affects all borrowers, particularly if there is feedback in demand across industries.

Moreover, there is cross-sectional evidence, among firms rather than industries, of relatively severe reactions to financial turmoil by firms with higher finance costs during the Depression, and evidence that a substantial number of firms faced very high costs of external finance by the mid-1930s. Kimmel (1939) found refusal or restriction of bank credit to manufacturing firms normally dependent on banks occurred for only 3.2 percent of the largest firms, but the refusal/restriction rate increased monotonically by size categories to a 30.2 percent refusal/restriction rate for the very smallest firms. Following Temin’s suggestion, one can compare these restriction/refusal rates to those in the recession of 1960. The Small Business Administration circulated a questionnaire asking firms, among other things, whether they received as much long-term credit as they requested from lenders. None of the 15 firms with asset values greater than $5 million reported being constrained, while 14 percent of 111 firms with lower asset values reported receiving less long-term credit than requested (Carson, 1963, p. 114). While these surveys may not be perfectly comparable, the results support Bernanke’s view that the Depression was a time of unusual credit hardship for small firms, even compared to other recessions.

In an interesting study that has received little attention, Hunter (1982) provides more detailed evidence along these lines. She groups firms by size categories and examines differences in firms’ balance sheet changes during the Depression. She argues that corporate liquidity preference increased substantially during the Depression, but increases in liquidity were confined mainly to large firms. Large firms were the only ones capable of improving their liquidity positions because of their superior access to financial markets. Hunter also compares the relative liquidity positions of small firms to large firms during the
Depression with that of other periods and finds that the financial stringency of
the Depression was associated with a uniquely large difference between the
liquidity positions of large and small firms. This evidence provides more direct
support for Bernanke’s position using standards of comparison suggested by
Temin (1989).

Calomiris and Hubbard (1991) also provide evidence supporting the view
that the mid-1930s were a time when many firms faced high costs of external
finance. Using firm-level data on dividend responses to the undistributed
profits tax in 1936 and 1937, corroborated by data on costs of securities issuing,
they find substantial heterogeneity in the costs of finance across firms, and that
a large number of firms (roughly a quarter of all firms paying taxes) had a
shadow price differential between internally generated and externally obtained
funds in excess of 20 percent. The investment of firms with high external
finance costs was highly sensitive to internally generated funds, while other
firms’ investment was not.

In summary, cross-sectional evidence confirms Bernanke’s (1983) interpre-
tation of time series patterns. The costs of external finance were quite high
during the Depression, these costs were particularly high for small, growing
enterprises. High finance costs reflected both the reduced creditworthiness of
firms as well as a contraction in the availability of “inside” bank debt.

Conclusion

The study of financial factors during the Great Depression has seen much
progress over the last 30 years. Friedman and Schwartz (1963) provided a
strong foundation of facts and provocative interpretations on which subsequent
research has built. This body of research also has shaped the way economists
think about monetary policy and financial markets more generally. For exam-
ple, many macroeconomists now believe that a large fraction of macroeconomic
disturbances have long lives, that monetary policy operates largely through its
effect of the real supply of bank credit, and that price flexibility can be
disruptive if it takes the form of unanticipated deflation. Those views are
compatible with, if not caused by, the last decade of research on financial
factors during the Depression.

In addition, some insights from the study of the Great Depression seem to
have methodological significance for the study of macroeconomic fluctuations,
not all of which has been absorbed by macroeconomics.

First, the literature on financial factors emphasizes the importance of
modeling the economy as an historical process. Prior experience governs the
information available to private agents and policy-makers, the stubbornness of
expectations, the balance sheets of agents, and the vulnerability of the economy
to disturbances. The “deep parameters” of the economy (which depend on
agents' information) are always changing. The reaction of the economy to exogenous shocks depends on interactions between shocks and time-varying state variables. Much progress in understanding financial factors during the Depression over the past 30 years has come from a willingness to take a careful look at specific moments of time and place them in an historical context, rather than collapse history into a long time series of aggregate data.

Second, recent research on financial factors in the Depression also has shown that panel and cross-section studies, which provide comparisons across countries and firms, can offer more insight than the repeated torture of the same domestic time-series aggregates, and such analysis does not depend on implausible assumptions of stationarity. The desirability of a greater emphasis on panel data to answer macroeconomic questions is further suggested by available empirical evidence that financing costs vary importantly across firms. As Schumpeter emphasized, business cycles are often driven by the activities of certain classes of firms and industries. The representative agent or firm approach to understanding macroeconomics is liable to leave key actors out of the play.

While financial factors can explain persistent reductions in the efficiency of capital allocations and economic activity, one must combine financial influences with other factors to explain protracted underutilization of resources (that is, unemployment and excess capacity). Here too recent research has shown the usefulness of abandoning the representative agent or firm assumption (Bernstein, 1987; Bresnahan and Raff, 1991, 1992; Margo, 1991, 1992; Wallis, 1989). These authors have emphasized that the disruption of the early years of the Depression brought endogenous responses in technological choice, the composition of consumption demand, and demands for labor skills which had important effects on aggregate production, capacity utilization, and employment. Like the new literature on financial factors, this body of research offers insights into how disturbances had long-lived effects on economic activity during the 1930s through their influences on the long-run survival of different firms and different technologies. For example, Bresnahan and Raff (1991, 1992) argue that cyclical decline ("shake-out") hastened technological change in the automobile industry. Financial factors are unlikely to provide the entire explanation for how early adverse shocks were transformed into the Great Depression.

A final lesson from the Depression for modern macroeconomics is the peril of assuming that shocks to technology (including changes in the cost of credit intermediation) are independent of shocks to monetary policy or other influences on "aggregate demand." Some recent macroeconomic studies of the post-World War II period assume independence of aggregate demand and supply shocks, and assume that aggregate-demand shocks have transitory effects while aggregate-supply shocks have permanent effects. These identifying restrictions permit one to measure the relative importance of aggregate supply and demand shocks for causing variation in GNP (Blanchard and Quah, 1988;
Shapiro and Watson, 1988). This approach results in large measured contributions to output variance at business cycle frequencies from supply shocks, which often are identified with exogenous technological change.

But this separation of aggregate-supply and aggregate-demand innovations is hard to justify from the perspective of the history of the Great Depression. A common emphasis of recent research on the Depression, including but not limited to work on financial factors, is that disturbances to aggregate demand like money-supply shocks may have had persistent effects on output through endogenous changes in firms’ and consumers’ balance sheets, technological shake-out, and endogenous changes in demands for different types of labor and consumption goods. In other words, monetary and other demand shocks had persistent effects on output, excess capacity, and unemployment through various channels connecting them to changes in the underlying structure of the economy.

The author is grateful to Michael Bordo, Barry Eichengreen, Michel Habib, Joseph Haubrich, Glenn Hubbard, Charles Kindleberger, Allan Meltzer, Frederic Mishkin, Jay Ritter, Hugh Rockoff, Joseph Stiglitz, Timothy Taylor, Peter Temin, David Wheelock, Eugene White, and Elmus Wicker for helpful comments on an earlier draft.

12 Of course, the notions that shocks to credit markets have a long-run impact on output, that technological change is endogenous to demand shocks, and that recessions are periods of reallocation of capital among firms that reflect heterogeneity with respect to technological opportunities and vulnerability to credit supply shocks are not new to macroeconomics. These were the basic building blocks of Schumpeter’s (1939) theory of business cycles.
References


Haubrich, Joseph G., "Nonmonetary Effects of Financial Crises: Lessons from the


Temen, Peter, Did Monetary Forces Cause the Great Depression? New York: W. W. Norton, 1976.


