Macroeconomic Principles and Monetary Policy

J. Alfred Broaddus, Jr.

It's a great pleasure and honor for me to be invited to participate in this Forum, although I have to tell you that I was more than a little intimidated when I learned that I would be part of a panel featuring Bob King and Tom Sargent. I take some comfort, however, from what Mike Dotsey told me when he first contacted me about this program seven or eight months ago. He said the panel would focus on optimal monetary policy, but he wasn't expecting me to provide a highly technical analysis, or even a low-tech analysis. Instead, he wanted me to talk about how I, as one fairly senior Fed monetary policymaker, use economic analysis and principles to arrive at policy positions and then present and defend them. This I think I can do, although I still feel a little uneasy with Bob and Tom so close at hand.

The first thing I need to say is that I *do* try to base my policy positions on solid economic analyses, as do my FOMC colleagues. And throughout my 11-year tenure as Richmond Fed president I've been blessed with exceptional policy advisors and a strong research staff who've made this possible. My principal policy advisor, Marvin Goodfriend, is well known to all of you, I'm sure. Our research director, Jeff Lacker, Bob Hetzel, and several other members of our staff provide strong support. Mike Dotsey was an important part of our policy team before the Philadelphia Fed got lucky and he moved up here. Finally, we've developed long-term advisory relationships with several leading university economists, most notably, Bob King and Ben McCallum. All of these people have helped keep me reasonably abreast of ongoing research in monetary economics, and, appropriately, they've insisted—some

This article is the text of an address given by J. Alfred Broaddus, Jr., president of the Federal Reserve Bank of Richmond, before the Philadelphia Fed Policy Forum on Managing the Recovery in Uncertain Times, at the Federal Reserve Bank of Philadelphia, on November 14, 2003. The author thanks his colleague, Marvin Goodfriend, for his assistance in preparing these remarks. The views expressed here are the author's and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

more vociferously than others—that I take advantage of what I've learned from them in formulating my policy positions. I've been happy to try to oblige.

Every once in a while I've heard people—including people close to and knowledgeable about the monetary policymaking process—opine that economic principles are not terribly important in the practical, day-to-day conduct of policy. Sure, they'll acknowledge, it's nice for central banks to support economic research related to policy and for staff economists to summarize the latest academic thinking for policymakers so that policymakers can participate effectively in policy discussions and debates with academic economists, the press, and others. But when the chips are down, the argument goes, and the FOMC sets the target level for the federal funds rate, the decision comes down to two things: (1) assessing what the latest economic and financial information says about the current condition of the economy and its prospects, and (2) determining how promptly and how strongly to respond to this information. Moreover, these judgments are strongly influenced by where the major economic indicators are expected to be in the period ahead in relation to their ranges in the past, and they are made with a generous amount of instinct and common sense.

Now there is more than a little truth in this characterization. I've been attending FOMC meetings at least part of the time since 1973, and I have certainly heard this view expressed in one way or another from time-to-time in the Committee's deliberations. Indeed, I've probably made comments like this myself.

It would be inaccurate and misleading, however, to suggest that this attitude has been a dominant one in the Committee either currently or in the past. On the contrary, economic analysis—including relatively recent developments in the professional economics literature—has frequently played a central role in determining policy, especially over the longer run. Unquestionably one of the Fed's greatest achievements over the last three decades was our role in, first, breaking the high inflation of the late 1970s and early '80s and, subsequently, helping bring the rate down to its current quite low level. The view that there was no exploitable systematic tradeoff between inflation and unemployment, which was gaining ground in the profession throughout the 70s, paved the way for this accomplishment. And the quantity theory corollary that central banks could control inflation by controlling money growth was its foundation.

Probably the best way I can describe how I use economics as a policymaker is to provide a few concrete examples drawn from my personal participation in FOMC meetings. (This may seem a bit self-centered, and I apologize if it does, but I think this is the best way for me to make the points I want to make.) As most of you no doubt know, full transcripts of FOMC meetings

are released to the public five years after a meeting.[1] Consequently, meeting transcripts are currently available through the meeting held December 16, 1997, which covers the first 40 meetings I attended as Richmond Fed president. I've reviewed these transcripts and selected three examples of how economic analysis guided my own thinking. The first involves my role in a "debate" regarding inflation targeting at the January 31–February 1, 1995, meeting. The second concerns my argument a few months earlier at the November 15, 1994, meeting that in principle the Fed should disengage as far as possible from foreign exchange market intervention. The final example comes from the May 20, 1997, meeting when I argued that an increase in trend productivity growth has important implications for interest rate policy not recognized by the macroeconomic models we typically use for monetary policy analysis.

In each case, I will describe briefly the context in which the policy issue came up and discuss the macroeconomic principle that guided my approach to the issue in question. Then, using the transcripts of the relevant FOMC meeting, I will describe how I used the principle as a basis for a policy recommendation. As always, the views that follow are my own and not necessarily those of any of my FOMC colleagues. This is, of course, a standard disclaimer that FOMC participants routinely recite. In this case, I have proof that my views are not necessarily those of my colleagues. Even a cursory reading of the relevant transcripts will make that abundantly clear.

1. INFLATION TARGETING (JANUARY 31–FEBRUARY 1, 1995, FOMC MEETING)

As you will remember, the first several quarters of the recovery from the 1990-91 recession were quite sluggish compared to most post-World War II recoveries up to that point. Real GDP grew at only a 2.6 percent annual rate from the trough in the first quarter of 1991 through mid-year 1993. Moreover, like the present recovery, it featured very weak growth in jobs, for which it also earned the sobriquet "jobless." In early 1994, however, the weakness in the economy began to abate, and the recovery gained momentum. By this time the CPI inflation rate had declined to 3 percent. To stimulate the recovery further, the Committee held the nominal funds rate at 3 percent for over a year. With inflation at 3 percent, the real funds rate was therefore zero. Most FOMC members[2] agreed that 3 percent inflation was not quite price stability, and probably everyone recognized that a zero real funds rate was inconsistent with containing inflation over the long run. Still, with the recovery only

[1] The transcripts are available on the Board of Governors website at www.federalreserve.gov/fomc/transcripts.

[2] Throughout this paper, references to FOMC members include non-voting as well as voting Reserve Bank Presidents.

beginning to accelerate as the year began, the strategy was to hold the line on inflation that year and then make the final step to price stability later.

The year 1994—my first as a voting member—turned out to be a moment of truth, or maybe I should say a year of truth, for the FOMC in the long fight for price stability. We were tested on two counts. First, there was an "inflation scare" in the bond market. The 30-year Treasury bond rate rose from a low of 5.9 percent in October 1993 to a peak of 8.2 percent in November 1994. Undoubtedly, a large portion of that increase reflected rising inflation expectations. Financial markets were far from confident that the Fed would succeed in containing inflation. Second, in February the Committee began to announce its funds rate target immediately after each FOMC meeting. This additional transparency meant that, henceforth, every interest rate action—or lack of action—would be scrutinized and second-guessed by the markets as never before.

In the event, we were able to raise the nominal funds rate by three percentage points between early 1994 and early 1995. And, since inflation held steady, the real funds rate rose by roughly the same amount over this period. The unemployment rate moved up following this tightening, but only slightly. Moreover, the long bond rate returned to about 6 percent, and people actually began to talk about the "death of inflation." It seems fairly clear in retrospect that our actions anchored inflation and inflation expectations. But, as we moved into 1995, I remember feeling that we'd been fortunate that we had accomplished this and that our credibility for low inflation was still not complete. The inflation scare in the bond market, in particular, made me think that we could still find ourselves in a position somewhere down the road where we would have to tighten policy sharply to shore up our credibility, with an attendant risk of setting off a recession.

It was in this context that I began to speak in FOMC meetings in favor of an inflation target. The initial discussions eventually led Chairman Greenspan to ask Governor Janet Yellen and me to lead a "debate" on inflation targeting at the January 1995 meeting. Janet spoke in opposition; I spoke in favor. The analytical principle that conditioned my support for targeting—rooted in the idea of rational expectations and reinforced strongly by discussions with Marvin and Bob Hetzel—was that by announcing an explicit long-run inflation objective, the FOMC would enhance the credibility of its commitment to low inflation and thereby reduce the risk that inflation would reaccelerate and, should it do so, reduce the cost of bringing it back down. In particular, I argued that anchoring inflation expectations more strongly with an explicit inflation objective would allow the FOMC to act more aggressively to help stabilize the economy in the short run, since with an explicit inflation anchor the Committee would be less concerned that such actions would reduce credibility and generate further inflation scares. In this environment, interest rate increases needed to hold the line on inflation would be less likely to cause recessions;

conversely, deep cuts in interest rates needed to stabilize the economy in a recession would be less likely to set off an inflation scare.

Let me fast-forward for a moment to the present. Inflation targeting has been receiving renewed attention recently. In the 1995 "debate," for a variety of reasons, I was willing to settle for an inflation objective that didn't necessarily include a numerical target. I felt that an FOMC commitment to the language of the proposed Neal Amendment,[3] for example, would suffice to capture the benefits I've just outlined. Today, however, with price stability achieved, I think a numerical range is definitely preferable. Specifically, our recent experience with disinflation and the proximity of the zero bound on the funds rate has convinced me that there is little to be gained—and considerable downside risk—in allowing trend inflation to drop below 1 percent. But if a lower inflation bound is warranted, then obviously (at least in my opinion) there should be an upper bound as well. For me, a 1 to 2 percent inflation target range for the core PCE would be acceptable.

I recognize that introducing an explicit inflation target would raise questions regarding exactly what its operational role would be in implementing policy. I am confident, though, that these issues could be addressed without unduly constraining the FOMC's traditional short-term stabilization policies. As I said in the 1995 debate, an inflation target "would not prevent the Fed from taking the kinds of policy actions that we take today to stabilize employment and output. What it *would* do (emphasis added) is to discipline us to justify our short-term actions designed to stabilize output and employment against our commitment to protect the purchasing power of our currency." I stand by that summary of the promise of inflation targeting.

2. INTERVENTION IN FOREIGN EXCHANGE MARKETS (NOVEMBER 15, 1994, FOMC MEETING)

My second example involves the viability of Federal Reserve participation with the U.S. Treasury in intervention operations in foreign exchange markets aimed at affecting the value of the U.S. dollar in these markets. A fundamental principle here, of course, is that intervention cannot have a sustained effect on the value of the dollar unless it is supported by basic monetary policy. Therefore, a problem arises immediately if the policy required to support a particular external objective for the dollar is inconsistent with the policy required to achieve broader domestic economic objectives. Beyond this, however, as I'll indicate in a moment, intervention can pose problems even where

[3] Representative Steve Neal of North Carolina proposed Amendments to the Federal Reserve Act in 1989, 1991, and 1993 that would have established price stability as the principal objective of Federal Reserve monetary policy. In the latter two years, the Amendment would have defined price stability as a condition where "the expected rate of change of the general price level ceases to be a factor in individual and business decisionmaking."

there are no direct conflicts between the policies required to support domestic and external objectives.[4]

In 1994 the Treasury and the Fed intervened frequently and visibly, often in conjunction with foreign central banks. These actions provoked an extended discussion of the Fed's participation in these operations at the November 1994 FOMC meeting. As the transcripts indicate, there was considerable disagreement among Committee members regarding the relative benefits and costs of this participation.

The comments I made in this discussion were guided by the principle that the Fed's credibility for low inflation is the foundation of effective monetary policy, and that public confidence in the Fed's independence in conducting monetary policy is the foundation of that credibility. Our experience over the preceding 15 years or more had made clear how difficult it is for the Fed to establish and maintain credibility. Consequently, I reasoned that we shouldn't allow anything to risk compromising our credibility.

Intervention, it seemed to me, did precisely that. The Fed is clearly the junior partner with the Treasury in foreign exchange intervention. To be sure, as a mechanical matter the Fed can follow the Treasury's lead in intervention operations without compromising its monetary policy independence by neutralizing the effect of its intervention actions on the funds rate through offsetting open market operations. There is little evidence, however, that such "sterilized" interventions can have a sustained effect on the exchange rate unless they are seen as signals of unsterilized policy actions in the future. Consequently, Fed participation in foreign exchange intervention with the Treasury risks creating doubt regarding whether monetary policy will support domestic or external objectives, and this confusion can undermine the credibility of the Fed's commitment to low inflation. I made this case in the November 1994 FOMC discussion. I also reminded the Committee of the high-profile, multination intervention in June of that year that was widely regarded in the press (including even non-national newspapers like the Richmond paper) as a failure. I argued that this kind of harshly negative publicity—even in a case, like this one, where the policy implications of the domestic and external objectives were not in direct conflict—could harm the Fed's credibility by creating an impression that the Fed was either unable or unwilling to achieve its policy goals more generally.

In sum, reasoning in this way, I concluded that the Fed had little, if anything, to gain and much to lose from participating in foreign exchange market interventions and that doing so would reduce the effectiveness of monetary policy over time. I therefore recommended at the November 1994 meeting that the Fed consider withdrawing from these operations, if not immediately,

[4] For a review of the issues surrounding foreign exchange market intervention, see Broaddus and Goodfriend (1996).

then gradually but persistently in some way. The meeting transcript shows that, while there was little support for my proposal to disengage, there was considerable sympathy with the logic of my argument and the economic rationale underlying it. Since that meeting, the FOMC has not formally changed its policy regarding intervention. But both the Treasury and the Fed have refrained from intervening in recent years. Circumstances have no doubt played a large role in this apparent reduction in the inclination to intervene, and I would certainly not claim that my statements in the FOMC meeting played any significant role in bringing this about. Whatever the reason for the change, however, the absence of these operations lately is clearly consistent with what economic analysis tells us about how to conduct monetary policy effectively.

3. INTEREST RATE POLICY AND HIGHER TREND PRODUCTIVITY GROWTH (MAY 20, 1997, FOMC MEETING)

From 1986 to 1990, non-farm business productivity grew only about 1.0 percent per year on average, which reflected the sustained slowdown in productivity growth that began in the mid-1970s. Trend productivity growth rose dramatically, however, in the 1990s; in fact, it tripled to an average of around 2.4 percent annually in the second half of that decade.

In 1996 and 1997, the FOMC began to recognize, along with other economic observers, the possibility that *trend* productivity growth might be undergoing a sustained increase. Economists understood that higher productivity growth would hold down inflation because it would take time for real wages to catch up. Unit labor costs would rise more slowly than the prices of final goods and services for a time and put downward pressure on inflation, as firms passed lower costs through to lower prices. Indeed, inflation hardly budged during the long boom in the late 1990s, even though labor markets tightened considerably. Rising trend productivity growth and the Fed's credibility for low inflation that I discussed earlier probably account to a considerable extent for the favorable inflation performance.

The implications of these developments seemed obvious. As long as rising productivity growth kept inflation low, the FOMC could refrain from raising its funds rate target. This was the generally held view when at the May 1997 FOMC meeting I brought up another channel, in addition to the unit labor cost channel, through which higher trend productivity growth might affect the choice of an appropriate funds rate target. I was motivated to do so by the possibility that trend productivity growth might have accelerated, which, as I just said, was beginning to be contemplated by the Committee.

In my economic statement at that meeting, I outlined this other channel as follows. I assumed that markets were confident that the Fed would hold the line on inflation so that inflation and inflation expectations would be stable.

How, in this situation, would higher trend productivity growth affect financial markets and real interest rates? Broadly, as I saw it, the improved productivity trend would cause firms to expect higher future earnings and workers to expect higher future wages. The point I emphasized was that at the then prevailing level of *real* interest rates, households and businesses would want to bring some of that expected increase in future income forward to the present. Workers might want to fix up their homes; firms might want to invest in new plant and equipment; and both households and businesses would try to finance such expenditures by borrowing against the expected future increases in income. Because the economy would not yet be producing this higher future income, however, real interest rates have to rise in order to prevent excessive *current* demand for goods and services from emerging. In other words, higher real interest rates would be required in order to raise the prices of goods and services consumed currently in terms of goods and services foregone in the future so that households and firms would be content to wait until the economy had actually produced the higher expected future output before trying to consume or invest it. The point was that, even if trend productivity growth were rising, and this increase reduced the inflation risk, real interest rates *still* needed to rise to prevent an unsustainable, credit-driven increase in aggregate demand that could lead to an unsustainable real boom.

My argument got no response during the FOMC discussion, although subsequently several Committee members expressed interest in it. In retrospect, though, I think the point looks pretty good. With the benefit of hindsight, the Committee might have done well to raise the funds rate target a little sooner than it did during the late 1990s boom. A somewhat more preemptive tightening of policy might have prevented some of the excess investment during the boom, and therefore the resulting weak investment that helped generate the recession and—until recently at least—slow the subsequent recovery.

As I indicated at the outset, my assignment today is to illustrate how economic analysis conditions my thinking about policy. In this particular case, the analytical result I just summarized, and that I used in the FOMC discussion, came from the "New Neoclassical Synthesis" macromodel we use at the Richmond Fed to think about monetary policy.[5] This "NNS" model has a real business cycle core that integrates growth and fluctuations, and it also has sticky prices that allow monetary policy to play a role in stabilizing inflation and employment. In this framework, it's easy to see the implications of an increase in trend productivity growth for interest rate policy. In particular, one can consider two NNS economies, where both have stable prices and full employment, and where consumption, investment, and output are all growing at the same rate as productivity. The only difference is that in one economy

[5] See Goodfriend (2002).

productivity is growing more rapidly than the other. The model shows that, in balanced growth equilibrium, the faster growing economy must have a higher real interest rate. If the central bank in this economy does not recognize this and holds real short rates below the equilibrium rate, borrowing and spending will exceed potential output in the short run and create an unsustainable boom in consumption, investment, and employment. The model cannot predict exactly how the boom will collapse if the central bank holds short-term rates too low for too long. It may end with accelerating inflation. Alternatively, where—as in the current cycle—the Fed has credibility for low inflation, it could end in recession accompanied by disinflation.

4. CONCLUSION

I hope these examples have illustrated reasonably clearly how at least one policymaker has used economic analysis in developing and arguing monetary policy positions in recent years. In particular, I hope the examples have suggested the scope of the opportunity for modern analytical tools to improve policy. Most importantly, I hope this discussion has helped underline the point I made at the outset: that while carefully monitoring incoming data and the evolution of the near-term outlook for the economy is an essential component of successful policymaking, it absolutely must be accompanied by solid economic analysis based on high quality research if monetary policy is to be as effective as it can be. I believe this need is well understood by my FOMC colleagues, and while this recognition may not have produced optimal monetary policy, I think it's definitely improved policy over the last two decades.

REFERENCES

Broaddus, J. Alfred, Jr., and Marvin Goodfriend. 1996. "Foreign Exchange Operations and the Federal Reserve." Federal Reserve Bank of Richmond *Economic Quarterly* 81 (Winter): 1–19.

Goodfriend, Marvin. 2002. "Monetary Policy in the New Neoclassical Synthesis: A Primer." *International Finance* 5 (2): 165–91.

Classical Deflation Theory

Thomas M. Humphrey

Deflation, the opposite of inflation, is a situation of falling general prices. It should not be confused with disinflation, which refers to a declining inflation rate that nevertheless remains positive. It was the successful U.S. disinflation of the 1990s, a disinflation that lowered the inflation rate sufficiently to create concern that further downward pressure might push it into negative territory, that spurred recent fears of deflation. These fears have materialized in Japan, where deflation coincides with cyclical recession and stagnant growth. Most famously, deflationary fears became reality in the 1929–1933 Great Contraction in the United States when prices fell by a fourth while output was falling by two-fifths.

Such episodes indicate that dread of deflation stems from its association with unemployment, business failures, and financial stress. Deflation tends to occur in cyclical slumps when collapses in aggregate spending force producers to cut prices continuously in a desperate effort to attract buyers. While these cuts eventually help to revive economic activity, they hardly work instantaneously. In the meantime, output and employment languish. The best alternative, therefore, may be to avoid deflation altogether by deploying monetary and fiscal policies sufficient to maintain economic activity at full capacity levels with low and stable inflation.

Absent in much of the recent worry over falling prices is the recognition that deflation is hardly a new topic or a new event. Classical (circa 1750–1870) monetary theorists, in particular, had much to say about it. Classicals, of course, abhorred deflation because, when unanticipated, it occasioned arbitrary and unjust redistributions of income and wealth from debtors to creditors. But classicals looked beyond these distributional outcomes involving equal but opposite transfers from losers to gainers to deflation's adverse effects on output and employment. As we will see, classicals attributed such adverse effects to

E-mail: Tom.Humphrey@rich.frb.org. For valuable comments, I am indebted to Margarida Duarte, Andreas Hornstein, John Walter, and John Weinberg. The views expressed in this article do not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

price-wage stickiness; to rising real debt, tax, and cost burdens owing to lags in the adjustment of nominal values of those variables to falling prices; to the hoarding (rather than spending) of cash in anticipation of future deflationary rises in the purchasing power of money; and to other determinants. In general, classicals assumed that deflation was unanticipated, the exception being their analysis of hoarding where they took expectations into account.

Generally, classicals wrote during or following periods of wartime inflation under inconvertible paper currencies. At such times the government had committed itself to return to gold convertibility at the pre-war parity. Such restoration, of course, meant that the price of gold, goods, and foreign exchange—all of which had risen roughly in the same proportion during the war[1]—had to fall to their pre-inflation levels. Achieving these price falls, however, required contractions of the money stock and so the level of aggregate nominal spending. Owing to the above-mentioned temporary rigidities in either final product prices or nominal costs of production, these falls in spending would depress output and employment first before they lowered costs and/or prices. With prices sticky, falling expenditure would show up in reductions in the quantity of goods sold. Unsold goods, the difference between production and sales, would pile up in inventories, thus inducing producers to cut back output and lay off workers. And if rigidities lodged in costs instead of prices, a reduction in spending would drive product prices below (inflexible) costs. The resulting losses (negative profits) would force producers to contract their operations. Eventually, however, rigidities would vanish, and prices and costs would fall in proportion to the monetary contraction. When this happened, real activity would return to its natural equilibrium or full-employment level, but not before workers and producers had suffered painful losses of income and employment.

Classicals analyzed these phenomena with a conceptual framework consisting of the quantity theory of money and the assumption of sticky product and/or factor prices. The quantity theory located the source of deflation in contractionary monetary shocks. And the sticky-price assumption explained the temporary adverse output and employment effects of the shocks. Together, these two pillars of the classical model reconciled the short-run nonneutrality with the long-run neutrality of money. In sum, on shocks and their propagation

[1] Classicals reasoned that the long-run equilibrium prices of goods, gold, and foreign exchange rose proportionally through the following causal chain. Inconvertible paper money, via its impact on total spending, determines domestic prices. Domestic prices, given foreign prices, determine the exchange rate so as to equalize, worldwide, the common currency price of goods. The exchange rate between inconvertible paper and gold standard currencies determines the paper price of bullion so as to equalize everywhere the gold price of goods. In symbols, $P = kM$, $P = EP^*$, and $G = EG^*$, where P denotes goods prices, k a constant, M the money stock, E the exchange rate, G the price of gold, and the asterisk* distinguishes foreign-country variables from home-country ones. Normalizing foreign-country variables at 1, converting the expressions into logarithmic form, and taking their derivatives yields $m = p = e = g$, where the lower-case letters represent proportionate rates of change of their upper-case counterparts.

through the economy's impulse-response mechanism, the classicals largely were in agreement.

Agreement did not extend to policies, however. Far from it. The classicals' quantity theory framework told them that the way to avoid deflation was to refrain from monetary contraction. But support of or opposition to that prescription varied with the policy objectives of the individual classical economist. Those preferring full employment at any cost endorsed the prescription even though it implied accepting the high prices established during the preceding inflation. Should the prescription conflict with the gold standard, so much the worse for the latter. Full-employment proponents were prepared to abandon metallic standards for a well-managed fiat paper standard and flexible exchange rates. On the other hand, those who favored restoring gold convertibility at the pre-war par were willing to countenance money-stock contraction, albeit at a gradual pace so as to minimize the costs of deflation.

The foregoing classical contributions have never been given their due recognition. To the best of this observer's knowledge, no systematic survey of classical deflation theory exists. Instead, one sees references to the neoclassical (circa 1870–1936) literature featuring contributions such as Irving Fisher's debt-deflation theory, his distinction between real and nominal interest rates, Knut Wicksell's notion of a painless fully-expected deflation, and Willard Thorp's empirical finding (see Laidler 1999, 187, 217, 223) of a relationship between secular deflation and the frequency, severity, and duration of cyclical depressions. According to Thorp, hard times were more likely to occur along a falling price trend than along a flat or gently rising one. While these and other concepts of the neoclassical literature are well known, the classical literature, by contrast, is largely ignored. This article is an attempt to repair this deficiency. It shows that the speculations of six leading classical monetary theorists—namely David Hume, Pehr Niclas Christiernin, Henry Thornton, David Ricardo, Thomas Attwood, and Robert Torrens—constitute a rich and coherent body of deflation theory, the constituent components of which survive today even as they are often wrongly attributed to neoclassical writers. To be sure, these six economists were not the only classicals to write on deflation. Nevertheless, they stand out as the seminal and influential ones. Their writings represent classical deflation theory at its best.

1. DAVID HUME (1711–1776)

Classical deflation theory begins with David Hume. Contrary to other classicals, he drew his inspiration not from the topical problem of the resumption of convertibility, but rather from an episode that occurred more than a hundred years before he wrote, namely the economic stagnation associated with the efflux of silver from Spain's colonies in the New World between 1560 and 1650. Hume's work is important because it established key features of the classical

theory. These included the assumptions that shocks are predominantly monetary, that deflation is partly unanticipated or unperceived owing to agents' lack of information on money supply variations, that prices lag behind prior changes in the money stock, and that monetary contractions therefore have nonneutral effects on real variables in the short-run if not the long. Most of all, his work demonstrates both the painfulness of deflation when prices are sticky and its painlessness when prices are flexible.

In his 1752 essay, "Of Money," Hume stresses the inertia of sluggish prices as the channel through which deflationary monetary contraction temporarily reduces output and employment. Sticky prices, which Hume attributes to the incomplete information price-setters possess on monetary changes and their resulting failure to act upon the changes, imply that deflationary pressure falls on real quantities first before it lowers prices. That is, from the equation of exchange $MV = PQ$ where M denotes the money stock, V its turnover velocity of circulation, P the price level, and Q the quantity of real output, it follows that with V constant (as Hume always assumed it to be) and P sticky, a fall in M must, by reducing aggregate demand MV, result in a corresponding temporary oversupply of goods that induces producers to cut output Q and lay off workers before they begin to lower prices. Money stock shrinkages, Hume wrote, "are not immediately attended with proportionable alterations in the price of commodities. There is always an interval before matters be adjusted to their new situation; and this interval is ... pernicious to industry, when gold and silver are diminishing ..." (Hume [1752] 1970, 40). Here is the source of the classical recognition of aggregate real effects of deflation, as opposed to purely distributional (creditor-debtor) effects.

Describing these pernicious real effects, Hume writes that "a nation, whose money decreases, is actually, at that time, weaker and more miserable than another nation, which possesses no more money, but is on the encreasing hand ....The workman has not the same employment from the manufacturer and merchant; though he pays the same price for everything in the market. The farmer cannot dispose of his corn and cattle; though he must pay the same rent to his landlord. The poverty, and beggary, and sloth, which must ensue, are easily foreseen" (40). Here is the source of the classical emphasis on the real costs of deflation.

In analyzing these deflationary costs, Hume distinguishes between those arising from one-time monetary contractions versus those stemming from continuous contractions. One-time contractions produce temporary losses occurring in the short run but not the long. At first, monetary shrinkage depresses real activity. Eventually, however, the real depression ends, and only lower prices remain. At this point Hume hints at a micro-foundations decision mechanism that the Swedish classical economist P. N. Christiernin later was to sketch in greater detail. Hume suggests that prices (and wages) start to fall only when price- and wage-setters, noticing that their inventories of unsold goods

and unused labor are abnormally high, interpret these excessive inventories as signaling the need for downward price-wage adjustment (Niehans 1990, 56). This correction continues until all perception errors are eliminated and real activity returns to its natural equilibrium level. Here is the source of the classical doctrine of the short-run nonneutrality and long-run neutrality of deflationary changes in the money stock.

Hume claimed that long-run neutrality holds for one-time but not for a steady succession of monetary contractions. The latter he believed to entail persistent real effects.[2] His explanation is straightforward (Humphrey 1982, 244–45). Continuous monetary contractions are partly unperceived and unadjusted for, perhaps because agents, lacking information, harbor static expectations and so expect the current money stock and price level to prevail in the future (Cesarano 1983, 198–99). Such surprise contractions forever stay a step ahead of sticky prices, perpetually frustrating their attempt to catch up. The result is that the lead of shriveling money over dwindling prices persists indefinitely, thus producing a permanent reduction in real activity. The upshot is that Hume, founder of the classical neutrality doctrine as it applies to levels of and one-time changes in the money stock, emerges as a believer in the long-run nonneutrality of continuous deflationary contractions of that stock. Because long-run nonneutrality holds for monetary expansions as well as for contractions, Hume's advice was to exploit the former while avoiding the latter. "The good policy of the magistrate," he said, "consists only in keeping it [the money stock], if possible, still encreasing; because, by that means, he keeps alive a spirit of industry in the nation, and encreases the stock of labour, in which consists all real power and riches" (Hume [1752] 1970, 39–40).

The preceding applies to the closed economy where the sticky-price assumption holds sway. Real effects vanish, however, when Hume—seeking now not to demonstrate nonneutrality but to banish mercantilist fears of a permanent loss of money—drops the assumption in his analysis of open trading economies. With prices now fully flexible, the specter of persistent deflation bringing losses in output and employment gives way to the notion of a specie-flow mechanism working swiftly to eliminate such phenomena.[3]

Let deflation occur: In his essay, "Of the Balance of Trade," Hume supposes that four-fifths of Britain's money stock is annihilated overnight with prices sinking immediately in proportion. At once British goods become cheaper on world markets and outsell all foreign goods at home and abroad.

---

[2] See the preceding quote where Hume refers to two nations possessing identical money stocks changing at different rates, negative in one country and positive in the other. Here is his belief that it is money's rate of change and not its quantity that matters for real variables in the long run.

[3] That Hume used his closed and open models for different purposes, the one to demonstrate short-run and possibly long-run nonneutrality, the other to expose mercantilist fallacies, perhaps accounts for his different treatments of price flexibility and inflexibility in the two cases.

The resulting export expansion and import shrinkage produces a trade balance surplus financed by inflows of monetary gold. The gold influx, by expanding the domestic money supply and hence total spending, bids up domestic prices to their pre-deflation levels. It all happens so fast that deflation is of extremely short duration. "In how little time," Hume asks rhetorically, "must this bring back the money which we had lost, and raise us to the level of all the neighbouring nations?" (62). Indeed, in other passages Hume suggests that quick-acting commodity arbitrage renders the process virtually instantaneous as it eradicates price differentials at home and abroad (Cesarano 1998; Niehans 1990, 56).

Hume's conclusion: Provided the world stock of monetary gold grows as fast as the real demand for it, deflation, at worst a transitory problem for open national economies, can never be a serious one. Deflation, in other words, is a speedily self-correcting phenomenon that brings its own remedy in the form of monetary expansion through the balance of payments. The result is to underscore the key importance of price inertia to Hume's analysis. Its presence in the closed case and absence in the open case renders deflation painful in the one and painless in the other.

2. PEHR NICLAS CHRISTIERNIN (1725–1799)

Hume was arguably the first classical economist to analyze deflation for countries on a metallic monetary standard and fixed exchange rates. Pehr Niclas Christiernin, a Swedish lecturer in economics at Uppsala University, was the first classical writer to do so for countries on a pure paper standard and floating exchange rates.[4] Sweden had converted to an inconvertible paper currency regime in 1745 and had suffered inflation under it during the Seven Years War (1755–1762). Christiernin wrote during the last years of that period when paper money expansion had raised the prices of goods, gold, and foreign exchange. With inflation reaching intolerable levels, Sweden's Parliament began to consider ways to arrest and reverse it. One group of politicians advocated deflation to restore prices to their pre-inflation levels.

Christiernin opposed such policies. In his 1761 *Lectures on the High Price of Foreign Exchange in Sweden* he argued that the best policy was to forgive the inflation that already had occurred and to stabilize prices at their prevailing, post-inflation level. The level of prices did not matter for real variables as much as changes in the level. Decisionmakers could get used to any price level, provided it was constant. They could not, however, tolerate the risks associated with deviations from that level. It followed that under no circumstances should prices be deflated. To do so when the "entire price

[4] On Christiernin, see Eagly (1971), Myhrman (1976), Niehans (1990, 56-59), and Persson and Siven (1993).

and wage structure" had become "fully adjusted to the current [depreciated] value" of the currency would be to "destroy our . . . prosperity" and plunge the economy into a slump (Christiernin [1761] 1971, 90). His fears were realized when in 1768 the policymakers, failing to heed his advice, deflated the level of prices by roughly fifty percent and precipitated a depression.

Echoes of Hume reverberate in Christiernin's claim that deflation diminishes "trade, industriousness, and the general welfare" (94). Like Hume, whose work he knew, Christiernin saw that price-wage stickiness—wage stickiness in excess of price stickiness being one of his innovations to Hume's analysis—transforms deflationary pressure into declines in output and employment. With prices and nominal wages slow to adjust (the latter more so than the former), monetary contraction leads to rising real wage costs, falling real profits, reduced real spending, and sinking real activity before it fully lowers money wages and prices. During the process, stagnation occurs in both the domestic and foreign trade sectors.

In the domestic sector, several influences besides Hume's sticky prices work to accentuate deflation's adverse real effects. According to Christiernin, these influences include undesired inventory accumulation (also mentioned by Hume), rising real debt and tax burdens, and deflationary expectations (the anticipated rate of return on holdings of money balances) that increase the demand for idle hoards of cash. All inhibit real spending, forcing it to fall short of its full-capacity level.

And in the export sector, a falling, or appreciating, exchange rate—a result of the money-induced contraction in income and demand including the demand for foreign exchange—combines with sticky prices to render the country's goods dear in terms of foreign currencies. This dearness reduces the foreign demand for, and consequently the domestic production of, the export goods. The appreciating exchange rate also makes foreign goods cheap in terms of domestic currency, thus shifting domestic demand toward imports and away from domestic import substitutes. By discouraging activity in the export and import-competing sectors, "a reduction in the price of foreign exchange . . . would have the worst possible consequence for commerce and industry throughout our nation" (89).

All this demonstrates that Christiernin did more than just build upon Hume's work. He advanced deflation theory markedly beyond Hume and took a giant step toward the modern analysis of the subject. Any compendium of elements that went into this step must include at least three of Christiernin's innovations. First, of course, is his restatement and refinement of Hume's sticky-price hypothesis, albeit with an asymmetric twist. "It is easy," Christiernin says, "for prices to adjust upward . . . but to get prices to fall has always been more difficult" (90).

Second are the explicit micro-foundations for price-wage stickiness barely hinted at by Hume. "No one," wrote Christiernin, "reduces the price of his

commodities or his labor until the lack of sales necessitates him to do so. Because of this [condition] workers must suffer want and the industriousness of wage earners must stop before the established market price can be reduced" (90). In other words, producers and workers lower their asking prices and wages only when unsold supplies of output and labor services materialize. Rising inventories of goods and labor constitute the signals that trigger reductions in prices and wages.

Third and most important are additional effects of deflationary monetary contraction beyond those adduced by Hume. These include (1) falls in the consumption and investment subcomponents of total spending, (2) undesired inventory accumulation mentioned above, (3) rising real burden of fixed, nominal lump-sum taxes, (4) rising real debt burdens and the associated rash of business bankruptcies, (5) growing deflationary expectations (the anticipated appreciation gains from holding money instead of goods) and the resulting increased demand for idle hoards of cash, (6) changes in the structure of relative prices, and finally (7) appreciating real exchange rates. An impressive list indeed.

Of the items on the list, Christiernin wrote the following: On monetary-induced falls in the separate consumption and investment components of spending: "A reduction of bank notes from circulation reduces everyone's consumption and the output of all sectors [including that of the capital goods sector]. The lack of capital [to equip labor and enhance its productivity] means unemployment and less industriousness among the working class, which results in less output" (93). On undesired inventories: a deflationary "shortage of money . . . causes[s] many goods to lie unsold (94), thus inhibiting production. On real tax burdens: nominal lump-sum "taxes . . . levied and paid in money . . . form a heavier burden . . . when . . . prices fall since more labor and goods are required to pay the same tax" (95).

Christiernin continues. Regarding real debt burdens and bankruptcies, he says, "When prices fall . . . the debtor must work longer and sell more commodities in order to retire his [fixed nominal] debt" (92). "[D]ebt . . . become[s] correspondingly more difficult to service and to repay. . . . Bankrupts . . . follow and the failure of one would pull down several more" (91). A debt-deflation spiral ensues as "all debtors . . . wish to sell all they had in order to pay off their debts before prices fell further" (94). Sellers hoping to beat the price fall flood the market with goods only to find that consumers "would not buy except at a low price—and even if they did buy (and the debts at the bank were repaid) the refunding of the principal to the bank would cause a new reduction in the circulation of money" (94). The result would be "nothing short of a complete credit breakdown" as "creditors [would] not dare loan their money for fear of debtors' inability to pay, and borrowers would not negotiate any loans because the fall in prices would [by reducing creditors' willingness to lend and so raising interest rates] mean they would have to pay more for less" (94-95).

A superior exposition of these phenomena, including the induced reduction of the money stock, would have to wait until the 1933 publication of Irving Fisher's debt-deflation theory of great depressions. Until then, Christiernin's version set the standard.

Likewise, his discussion of deflationary expectations and the resulting increased demand for idle balances was not surpassed until the 1920s. On these expectations-induced demands for cash, Christiernin wrote that "Deflation . . . increase[s] the need for money because of speculation and hoarding. When it was known that bank notes were becoming more and more valuable as a result of reductions in the money supply and that all prices in time would consequently fall, everyone would await that time and in the interim would not purchase more than the bare essentials" (94), but would hoard cash instead.

Finally, on the structure of relative prices, he says that "the price impact of a reduction in the money supply is not uniform: Not all prices fall; not all prices fall at the same time .... Prices only fall for those goods that are less essential or that are in over supply" (90). Christiernin never explained why these relative price changes, which by definition average out to zero, adversely affect aggregate activity. They could do so if producers, treating the variability of relative prices as a measure of business uncertainty and risk, interpret the changes as evidence of such increasing risk and accordlingly cut production. Of classical deflation theorists, only Thomas Attwood, a Birmingham banker, pamphleteer, and Member of Parliament, would provide a superior statement of the effects of the unevenness of price falls during deflation.

3. HENRY THORNTON (1760–1815)

Although Christiernin's work foreshadowed Thornton's, there is no reason to believe that Thornton, a London banker, Member of Parliament, and author of one of the nineteenth century's two best books on monetary theory, knew of it.[5] For one thing, Christiernin wrote in Swedish, a language inaccessible to Thornton and his English contemporaries. Then, too, Thornton differed from Christiernin on certain points. True, both men feared the effects of monetary contraction deliberately engineered to reverse a preceding rise in prices. But Christiernin always attributed the prior price rise solely to overissue of paper, whereas Thornton recognized that real shocks also could be a cause. And while Christiernin opposed reversing monetary overissues and the price rises caused by them, Thornton, at bottom a hard money man, favored such reversal, albeit at a cautious, moderate pace to avoid precipitating panics. Thornton reasoned that overissue of the paper component of the money stock could, unless reversed by policy, persist "permanently" even as gold—which in

[5] Much has been written on Thornton. Studies highlighting his deflation analysis include Hicks (1967) and Salerno (1980, 357–400).

Thornton's day still circulated as coin that could be melted down for export—was flowing out. That is, overissue could persist long enough to keep prices high and render the country's goods uncompetitive in world markets. It was price rises caused by real rather than monetary shocks whose reversal by deliberate monetary contraction he opposed. Real shocks, unlike overissues of paper money, tended to be temporary and self-reversing. That being so, it made no sense to put the economy through the wringer of monetary deflation to correct something that would quickly correct itself.

Thornton wrote during the first, or inflationary, phase of the famous Bank Restriction period (1797–1821) when the exigencies of the Napoleonic wars forced the Bank of England to suspend the convertibility of its notes into gold at a fixed price upon demand. The suspension of specie payments and the resulting move to an inconvertible paper regime was followed by rises in the prices of goods, gold, and foreign exchange. An influential group of classicals known as the strict bullionists arose to attribute these inflationary phenomena solely to the redundancy of money and to accuse the Bank of taking advantage of the absence of a convertibility constraint to overissue the currency. Against this purely monetary explanation Thornton contended that inflation must persist for several years before bullionist critics could claim they had proof enough to blame it on the Bank. For shorter periods, negative real shocks beyond the Bank's control might be the culprit.

The negative real shocks that concerned Thornton were nonmonetary disturbances to the balance of payments. These disturbances included domestic harvest failures as well as wars and the associated extraordinary foreign expenditures on subsidies to allies and on the maintenance of troops abroad. All tended to put the balance of payments into deficit.

Conventional wisdom at the time called for correcting such deficits with deflationary monetary contraction. Such contraction would, by making the country's goods cheaper both at home and abroad, spur exports, check imports, and so restore equilibrium in the external accounts. Thornton, however, opposed this remedy on the grounds that, by precipitating a depression and disrupting production, it would reduce the output of goods available both for export and for import substitution and so worsen, rather than correct, the payments deficit. In explaining this perverse outcome, Thornton, like Christiernin, identified channels additional to Hume's sticky price circuit through which deflationary contraction depresses real activity.

First was a money-demand channel. Unlike Christiernin, Thornton never stressed the influence of deflationary expectations on cash holdings. But he did note that manufacturers and merchants have a well-defined demand for money balances, balances held for the purpose of conducting transactions, paying suppliers, and compensating workers.

Given this money demand, a sudden, sharp contraction of the money stock creates a cash deficiency that depresses real activity. Merchants, attempting

to rebuild their balances to the desired level "delay making the accustomed purchases of the manufacturer . . . .[whose] *sales* . . . are, therefore, suspended" (Thornton [1802] 1939, 118, italics in original). This sales stoppage adversely affects manufacturing output and employment—all the more so as the manufacturer is at that same time being "pressed for a prompter payment than before" by his creditors while his continued outlay on labor and materials means that "his money is going out while no money is coming in" (118). Because of these considerations, the "manufacturer, on account of the unusual scarcity of money, may even, though the selling price of his article should be profitable, be absolutely compelled by necessity to slacken, if not suspend, his operations" (118). In short, merchant reluctance to buy transforms an excess demand for money into a deficient demand for manufactured goods.

Thornton's second channel through which deflation depresses real activity runs through sticky money wages. By refusing to fall when prices fall, sluggish nominal wages translate into rising real wages and falling real profits that destroy incentives for employment and production. Christiernin, of course, had said the same thing.

But Christiernin had said nothing about the source, or cause, of wage stickiness in the face of falling prices. Here Thornton had the edge. He located this source in workers' beliefs that under the gold standard then prevailing (though temporarily suspended for the duration of the Napoleonic wars) price falls are transitory and reversible. Workers, expecting deflated prices to return soon to traditional levels, naturally are unwilling to accept wage cuts in the interim. In his 1802 *Paper Credit of Great Britain,* Thornton expressed the whole matter in a passage that for clarity, precision, and perspicacity is unrivaled in the classical literature and is hardly surpassed today.

> [A] diminution in the *price* of manufactures . . . may also, if carried very far, produce a suspension of the labour of those who fabricate them. The masters naturally turn off their hands when they find their article selling exceedingly ill. It is true, that if we could suppose the diminution of bank paper to produce permanently a diminution in the value of all articles whatsoever, and a diminution, as it would then be fair that it should do, in the rate of wages also, the encouragement to future manufactures would be the same, though there would be a loss on the stock in hand. The tendency, however, of a very great and sudden reduction of the accustomed number of bank notes, is to create an *unusual* and *temporary* distress, and a fall of price arising from that distress. But a fall arising from temporary distress will be attended probably with no correspondent fall in the rate of wages; for the fall of price, and the distress, will be understood to be temporary, and the rate of wages, we know, is not so variable as the price of goods. There is reason, therefore, to fear that the unnatural and extraordinary low price arising from the sort of distress of which we now speak, would occasion much discouragement of the

> fabrication of manufactures (Thornton [1802] 1939, 118–19, italics in original).

Thornton's third channel features wastes and inefficiencies of capacity underutilization and resource misallocation. It is through this channel that a deflationary "diminution of notes prevents . . . industry . . . from being so productive as it would otherwise be" (119). He sketches a scenario in which deflation leads to the squandering of inputs as projects are halted and abandoned and the labor embodied in them is lost. Capital equipment is shut down only to produce nothing during its period of idleness. Unsold goods pile up in inventories where they lose value through physical deterioration and obsolescence. Then too, cash-starved producers, in a frenzy to obtain liquidity, dump specialized, unique goods on undifferentiated markets unsuited to their absorption. For all these reasons, Thornton says, "There cease . . . to be that regularity and exactness in proportioning and adapting the supply to the consumption, and that dispatch in bringing every article from the hands of the fabricator into actual use, which are some of the great means of rendering industry productive, and of adding to the general substance of a country" (120–21).

It follows that "Every great and sudden check given to paper credit not only operates as a check to industry, but leads also to much . . . misapplication of it" (121). This wastage spells a further reduction in national product, or as he puts it, a "diminution of the general property of the country . . . and, of course, a deduction also from that part of it which forms the stock for exportation" (121). In short, deflation impairs efficiency whose attenuation pushes output further below its full capacity potential.

Thornton concludes that deflation is the wrong way to spur exports and check imports and thus to remedy real-shock-induced deficits in the balance of payments. Deflation is ill-advised because "To inflict such a pressure on the mercantile world as necessarily causes an intermission of manufacturing labour is obviously not the way to increase that exportable produce, by the excess of which, above the imported articles, gold is to be brought into the country" (118). Better to ride out the real disturbances with unchanged, or even increased, issues of paper money until the disturbances correct themselves.

To summarize, Thornton's position on deflation was this: Deflate to reverse price rises emanating from monetary overissue. But never deflate to correct real shocks to the balance of payments. Such deflation operates through the channels described above to lower real output of exportable and import-competing goods and thus has a perverse effect on the trade balance. Deflation in this case is unnecessary anyway because real shocks are temporary, and the balance of payments will correct itself.

4. DAVID RICARDO (1772–1823)

Of classical monetary theorists, David Ricardo has the reputation of believing that money and price-level changes have no effect on aggregate real variables in either the short run or the long. But this reputation is not entirely warranted. His awareness of the real costs of deflation underlies his policy rules for restoring equality between the market and mint prices of gold after arbitrage-inhibiting inconvertibility has allowed the two to move apart. Upon resumption of convertibility, he would deflate away small, but not large, gaps between the two prices. Large gaps he would eliminate by raising the mint price to the prevailing market price instead of lowering the market to the old mint price. He also recommended gradualism in deflation and the avoidance of policy mistakes that worsen deflation.[6]

Ricardo was writing in the second, or deflationary, phase (1815–1821) of the Bank Restriction era when wartime inflation had given way to post-war deflation, and the authorities were considering how to implement resumption. During the war, the price of gold had undergone substantial inflationary upward drift such that bullion commanded a premium over its mint price. The decision to resume gold convertibility spelled the elimination of this premium. No such premium could exist when agents could convert paper, at the fixed mint price, into gold for resale on the market. Arbitrage would eradicate the premium. But the authorities could determine, through their setting of the mint parity, which price—market or mint—would adjust. Two-price equality could be achieved either through a lowering of gold's market price to the pre-war mint parity or, by devaluing (debasing) the standard, through a raising of the mint parity to gold's going market price. The first option involved painful deflationary monetary contraction. The second and far less painful option involved accepting the gold price rise that had occurred during the war, re-basing the mint parity at that price, and keeping the money stock unchanged.

Ricardo favored the deflationary option, but with two major provisos. First, the gap between market and mint prices of gold should not be too large. Deflation to eliminate a 5 percent gap was one thing, deflation to eliminate a 30 percent gap quite another. Should the gap be 30 percent or more, Ricardo was prepared to abandon restoration at the old parity for a new parity established at the prevailing price. Instead of deflating back to par, he would leave prices as they were. As he wrote in a September 1821 letter to John Wheatley, "I never should advise a government to restore a currency, which was depreciated 30 percent, to par; I should recommend . . . that the currency be fixed at the depreciated value by lowering the standard [i.e., raising the par], and that no further deviations take place. It was [a] currency . . . within 5 percent [of par]

[6] On Ricardo's view of the costs of deflation see Hollander (1979, 488–500) and Laidler (2000, 29–31).

and not with a currency depreciated 30 percent, that I advised a recurrence to the old standard" (Ricardo [1821] 1951, IX, 73–74).

Ricardo's advice is of more than antiquarian interest. It speaks to today's distinction between zero-inflation versus price-level targeting where the former allows for price drift and the latter does not. What should a central bank do when confronted with an upward drift in the price level? Should it accept such drift and thereafter stabilize the inflation rate about the new, higher price level? Or should it refuse to accommodate drift and instead deflate prices back to their old target level? Ricardo's position on these matters was clear. In the case of large market-minus-mint-price gaps, he was for accommodating drift. Instead of deflating gold prices back to par, he would leave them as they were. And in other passages, he made it clear that while he would accept the price drift that already had occurred, he would rely upon restored convertibility to prevent further drift from the re-based gold price level. So while he was not an inflation targeter strictly speaking, he at least was for resetting the price-level target when the old one implied excessive deflation.

Ricardo's second proviso was that deflation, once the Bank of England had decided to accomplish it, should be conducted gradually. Influenced by Thornton's work, Ricardo saw that precipitous deflation would wreck the economy. It "would be attended with the most disastrous consequences to the trade and commerce of the country, and . . . would occasion so much ruin and distress, that it would be highly inexpedient to have recourse to it as the means of restoring our currency to its just and equitable value" (Ricardo [1810–11] 1951, III, 94). Such sharp and sudden deflation should be shunned absolutely. Gradualism, not abruptness, was the key to conducting deflationary policy to close small price gaps. "If gradually done," he said, "little inconvenience would be felt" (94). In this connection he suggested transitory devaluation, that is, setting a temporary new mint parity to which the market price would conform, and then lowering both in easy stages to the old mint par.

Ricardo's recommendation of gradualism said nothing about rational expectations. As Lucas, Sargent and Wallace, and others have taught us, however, a fully expected deflation adjusted for in all contracts should have no real effects. Now the Bank's pre-announced return to gold at the old parity was a perfect example of a deflation that would seem to be fully expected. If so, the Bank could eschew gradualism and deflate to par immediately without fear of precipitating a recession.

In Ricardo's defense, however, as well as that of other classicals writing in the sticky-price tradition, it must be noted that with inflexible prices even perfectly foreseen monetary contractions can have negative output effects. If so, then price rigidity rather than neglected rational expectations does the damage. In any case, it was not until 100 years after Ricardo wrote that the Swedish economist Knut Wicksell explicitly enunciated the case for rational expectations under flexible prices and the corresponding case for immediate

movements to par. "If the price fall is clear beforehand and can be fairly forseen," said Wicksell, "businessmen ought to ... be in position to adapt themselves to the expected increase in the value of money ... so that they can work without ... losses" (Wicksell 1918, 1920, quoted in Boianovsky 1998, 225). With rational expectations and flexible prices rendering money-stock contraction neutral in its real effects, policymakers can deflate to par in one step. Here is the crucial ingredient missing in Ricardo's analysis.

Ricardo's greatest fear was that the Bank's own policy errors upon resumption would worsen deflation. In particular he feared that the Bank, in acquiring extra bullion reserves so that it could convert paper notes and deposits into gold coin when its customers so requested, would exert a heavy demand for gold in world markets. This demand would bid up gold's value or, in other words, lower the price of goods in terms of gold. Deflation of this price would augment deflation of the money price of gold to put double downward pressure on the money price of goods.

Ricardo's argument was quite ingenious. Recognizing that the money price of goods P is by definition equal to the multiplicative product of the money price of gold G and the world real gold price of goods R, or $P = GR$, he saw that deflation accompanying resumption could stem from two sources. Source number one was the fall in the gold premium, or money price of gold G, necessary to restore the market price of the metal to its mint parity. Source number two was a fall in the real gold price of goods R caused by the additional English demand for the fixed world supply of the metal consequent upon resumption.

To prevent the latter source of deflation, Ricardo, in his 1816 *Proposals for an Economical and Secure Currency*, offered his famous ingot plan in which the English money stock would consist solely of paper currency backed by, and convertible into, a reserve of bullion ingots. Such a paper currency offered the advantage (over gold coin) of greater flexibility in operation. It could, at the cost of variations in the reserve ratio, readily be expanded or contracted to accommodate temporary shifts in money demand.[7]

This, however, was but an incidental benefit of the scheme. Ricardo stressed the essential one: By abolishing gold coin and the Bank's need to hold specie reserves to accommodate increased requests for such coin, his ingot plan would minimize England's demand for the fixed world stock of bullion. True, there might be an export demand for gold ingots as well as a demand

[7] "Whenever merchants ... have a want of confidence ... more money ... is in demand; and the advantage of a paper circulation ... is, that this additional quantity can be presently supplied without occasioning any variation in the value of the whole currency ... whereas with a system of metallic currency, this additional quantity cannot be so readily supplied, and when it is finally supplied, the whole of the currency, as well as bullion, has acquired an increased value" (Ricardo [1816] 1951, IV, 58).

coming from the arts and industry. But such demands would be negligible in comparison with the (abolished) demand for coin and coin reserves.

By relieving the Bank of the need to hold large metallic reserves, the plan would largely remove that institution from world gold markets where it consequently would exert little downward pressure on the gold price of goods R. With gold's real value remaining largely unchanged, deflation of the general level of commodity prices P would be limited to the fall in specie's nominal price G necessary to restore parity.

Ricardo estimated that a deflation of no more than 5 percent would return gold to par under his ingot plan. But neither Parliament nor the Bank would adhere to his scheme. Indeed, Parliament set the plan aside before it could be executed, and the Bank filled its coffers with gold, which it had drained from the world market. The resulting deflation was twice what Ricardo estimated it would have been had the Bank been allowed to implement his plan. This entire excess deflation he blamed on policy mismanagement of the resumption.

To summarize, Ricardo's concern with falling prices is evident in the rules he laid down for dealing with deflation: Deflate only to eliminate small price deviations from target. In the case of large deviations, eschew deflation and accept price drift. If you must deflate, do it gradually. Avoid policy mistakes that worsen deflation.

5. THOMAS ATTWOOD (1783–1856)

Writing in the deflationary phase of the Bank Restriction period, Birmingham banker and pamphleteer Thomas Attwood, the most radical anti-deflationist of the classical era, would have none of Ricardo's proposals.[8] All of them, gradualism and devaluation included, envisioned stabilizing the market price of gold, if not at its pre-war parity, then at least below its wartime peak when full employment had prevailed. Ricardo's price-stabilizing objective was anathema to Attwood who, hailing from an area particularly hard hit by post-war depression, advocated full employment instead. "The first and most important duty for the Legislature to attend to," he said, "is to take care that an ample demand for labour is restored and maintained throughout the country" (Attwood [1843] 1964, 17). By ample demand, he meant "a demand for labour . . . permanently greater than its supply" (17).

For Attwood then, full employment was the overriding policy goal, and price increases were the essential means of securing it. Government had "the duty . . . to continue the depreciation of the currency until full employment is obtained and general prosperity" (Attwood [1831–32], 467, quoted in Corry 1962, 86). The policy authorities, upon reaching the employment target,

[8] On Attwood, see Fetter (1964, vii-xxviii), Laidler (2000, 25–27), O'Brien (1975, 164–65), and Viner (1937, 173, 186–87, 195, 199, 212–14, 289).

should permit prices to rise to levels compatible with it unconstrained by arbitrary ceilings. Inflation up to this height (but not beyond) was acceptable, even desirable. For when you "restore the depreciated state of the currency ... you restore the reward of industry, you restore confidence, you restore production, you restore consumption, you restore everything that constitutes the commercial prosperity of the nation" (Attwood [1819] 1964, 66). But deflation, the evil "which ought most to be guarded against, which produces want of employment, poverty, misery, and discontent in nations" (Attwood [1843] 1964, 18) must be avoided at all costs.

Fearing deflation even more than did his classical peers, Attwood saw it as harmful because it worked its way slowly, unevenly, haphazardly, and disruptively through the price structure. "If prices were to fall suddenly, and generally, and equally, in all things," he wrote, "*and if it was well understood that the amount of debts and obligations were to fall in the same proportion, at the same time*, it is possible that such a fall might take place without arresting consumption and production, and in that case it would neither be injurious or beneficial in any great degree, but when a fall of this kind takes place in an obscure and unknown way, first upon one article and then upon another, without any correspondent fall taking place upon debts and obligations, it has the effect of destroying all confidence in property, and all inducements to its production, or to the employment of labourers in any way" (Attwood [1817], 78–79, quoted in Viner 1937, 186, italics in original).

Equally important, deflation lowered product prices below wages and other contractually fixed costs. And when "the prices of commodities are suffered to fall ... within the level of the *fixed charges and expenses* ... the industry of the country dies" (Attwood [1826] 1962, 42, italics in original). It dies because profit margins, the difference between prices and costs, vanish and with them the means and the motive to produce. Output and employment then decline in a self-reinforcing downward spiral. For the same falling prices that combine with rigid cost elements to depress profits also prompt an unloading of stocks of goods. This dumping of goods puts further downward pressure on prices and profits causing still another unloading of stocks, etc. The downward movement continues until stocks are exhausted and the resulting shortage of goods spurs a rise in prices that ends the process at the trough of the cycle. This sequence brings great suffering to unemployed workers and hardship to businessmen. For these reasons, it is crucial that deflation be averted.

To Attwood the policy choices were clear: Use expansionary policy, inflating if necessary, to achieve full employment. Require or induce the Bank of England "to encrease the circulation of their notes ... until all the labourers in the kingdom are again in full employment at ample wages" (Attwood [1819] 1964, 44). Once the employment target is reached, accept the market price of gold that coexists with it and establish that price as the new mint parity. Never attempt to deflate away premia in the market over the mint price of gold, not

even when they produce external drains that threaten exhaustion of the nation's gold reserve. Instead, be prepared to abandon the gold standard with its system of fixed exchange rates for an inconvertible paper currency regime with floating rates. The latter regime gives the government the autonomy to pursue its full employment objectives free of external constraints.[9]

6. ROBERT TORRENS (1780–1864)

No survey of classical deflation theory would be complete without mention of Robert Torrens's efforts to incorporate tariffs into the theory. Already in 1812 he had recommended raising the domestic tariff as a means of preventing price declines when restoring convertibility. Admitting that such a restriction on trade would be to sacrifice the advantages of international specialization and division of labor, he argued that the avoided costs of deflation outweighed the forfeited gains from free trade (Viner 1937, 207).

Then in his 1844 *The Budget* he showed how the imposition of a foreign tariff could foist deflation on the home country in a gold standard regime. He established this result with the aid of a two-country, two-good model—his famous Cuba-England, sugar-cloth case.[10] His model has the export sector of each country specializing in the production of the good, fabricated at constant cost, in which it has the comparative advantage. The model also features unit elastic demands for both goods in both countries.

With these assumptions, Torrens showed that Cuba's imposition of a 100 percent import tariff on cloth creates a trade balance deficit in England. The resulting specie drain that finances the deficit causes England to lose one-third of her monetary gold stock to Cuba before the trade balance re-equilibrates itself. No country, Torrens thought, could endure a monetary contraction and proportional price deflation of that magnitude. The collapse of prices would bring ruinous rises in the real burden of debts, wages, and taxes the nominal values of which were sticky and responded sluggishly to deflationary pressure. Calamitous "crisis, . . . national bankruptcy, and revolution would be the probable results" (Torrens 1844, 37, quoted in Robbins 1958, 203).

To Torrens the policy implications were clear. Fight tariffs with tariffs. Cancel the deflationary effects of foreign duties by erecting compensating retaliatory duties at home. Such retaliatory duties, he said, "would bring back the metals . . . restore the circulation to its former amount, raise the price of all domestic products, and mitigate the pressure of the debt." (37, quoted

[9] "Self-existent, self-dependent, liable to no foreign nations, entirely under our own controul; contracting, expanding, or remaining fixed, according as the wants and exigencies of the community may require, a *non-convertible* Paper Currency presents every element of national security and happiness . . ." (Attwood [1826] 1964, 34, italics in original).

[10] On Torrens's Cuba case see O'Brien (1975, 191–94), Robbins (1954, 199–203), and Viner (1935, 298–99, 322, 463).

in Robbins, 203). In a word, practice reciprocity. Raise tariffs *pari passu* with the foreigner, and lower your tariff only if he lowers his. Needless to say, such reciprocity considerations did not sit well with Torrens's classical contemporaries, all of whom were unilateral free traders. But at least Torrens had highlighted a possible conflict between the goals of unilateral free trade and anti-deflationism in a tariff-ridden, gold standard world. For better or worse, Torrens's arguments still are employed today by those who put the blame for domestic deflationary pressures on foreign commercial policies.

7. CONCLUSION

If the classical writers surveyed in this essay are at all correct, then current concern over deflation is fully justified: For the essence of the classical doctrine is that there is every reason to spare the economy the adverse real effects of deflationary pressure. These effects, whether caused by lags of sticky prices behind money; or by lags of sluggish wages, interest, taxes, and other costs behind prices; or by rising real debt burdens and the resulting defaults and bankruptcies; or by cash hoarding in anticipation of future price declines; or to a combination of these and other causes, are likely to be painful in the extreme—especially so for deflations that are sharp, sudden, or sustained.

It follows that a policy of inflation targeting may be superior to price-level targeting as a means of eluding deflation. Suppose inflationary shocks and/or policy mistakes and inertia have caused or permitted prices to drift upward. Under inflation targeting, the central bank forgives the price drift that has occurred and thereafter stabilizes inflation about the new price level. It disinflates to its zero (or low positive) target rate of inflation at this price level but need not lower the price level itself. By contrast, under price-level targeting the central bank must engineer deflation and the recession it brings in order to lower prices to target.

Of course, deflation under certain circumstances might not be a bad thing, that is, might have no adverse real effects. If so, policymakers could ignore it or implement it with impunity. Such would be the case for deflations that (1) are always fully expected, (2) occur in a setting of complete wage-price flexibility, and (3) stem from productivity-induced growth in aggregate supply rather than monetary-induced contractions in aggregate demand. With the possible exception of Hume, however, classicals paid insufficient attention to these considerations and left their discovery to their neoclassical successors. But even if they had acknowledged them, they would have merely distinguished between bad and good (benign) deflations. As it was, they concentrated on harmful deflations. And it is these deflations that policymakers should seek to avoid. This lesson remains as valid today as it did in classical times.

REFERENCES

Attwood, Thomas. [1816] 1964. *The Remedy; or, Thoughts on the Present Distresses*. In *Selected Economic Writings of Thomas Attwood*. Ed. Frank W. Fetter. London: London School of Economics and Political Science, University of London.

_____. 1817. *Prosperity Restored; or Reflections on the Cause of the Present Distresses and on the Only Means of Relieving Them.* London.

_____. [1819] 1964. *A Letter to the Earl of Liverpool*. In *Selected Economic Writings of Thomas Attwood*. Ed. Frank W. Fetter. London: London School of Economics and Political Science, University of London.

_____. [1826] 1964. *The Late Prosperity, and the Present Adversity of the Country, Explained*. In *Selected Economic Writings of Thomas Attwood*. Ed. Frank W. Fetter. London: London School of Economics and Political Science, University of London.

_____. 1831–32. In *Report from the Committee on Secrecy in the Bank of England Charter: With the Minutes of Evidence*. Parliamentary Papers, Commons. (772), VI: 452–68.

_____. [1843] 1964. *Thomas Attwood's Letter to Sir Robert Peel on the Currency*. In *Selected Economic Writings of Thomas Attwood*. Ed. Frank W. Fetter. London: London School of Economics and Political Science, University of London.

Boianovski, Mauro. 1998. "Wicksell on Deflation in the Early 1920s." *History of Political Economy* 30 (2): 219–73.

Cesarano, Filippo. 1983. "The Rational Expectations Hypothesis in Retrospect." *American Economic Review* 73 (1): 198–203.

_____. 1998. "Hume's Specie-Flow Mechanism and Classical Monetary Theory: An Alternative Interpretation." *Journal of International Economics* 45: 173–86.

Christiernin, Pehr N. [1761] 1971. *Lectures on the High Price of Foreign Exchange in Sweden*. Ed., trans. R. Eagly in *The Swedish Bullionist Controversy*. Philadelphia: American Philosophical Society.

Corry, B. A. 1962. *Money, Saving and Investment in English Economics 1800-1850*. New York: St. Martin's Press.

Eagly, Robert V., ed., trans. 1971. *The Swedish Bullionist Controversy; P. N. Christiernin's "Lectures on the High Price of Foreign Exchange in Sweden"* (1761). Philadelphia: American Philosophical Society.

Fetter, Frank W. 1964. Introduction to *Selected Economic Writings of Thomas Attwood*. London: London School of Economics, University of London.

Hollander, Samuel. 1979. *The Economics of David Ricardo*. Toronto: University of Toronto Press.

Hicks, John R. 1967. "Thornton's Paper Credit." In *Critical Essays in Monetary Theory*. London: Oxford University Press.

Hume, David [1752] 1970. "Of Money" and "Of the Balance of Trade." In D. Hume's, *Writings on Economics*. Ed. E. Rotwein. Madison: University of Wisconsin Press.

Humphrey, Thomas M. 1982. "Of Hume, Thornton, the Quantity Theory, and the Phillips Curve." Federal Reserve Bank of Richmond *Economic Review* 68 (November/December): 13–18.

Laidler, David. 1999. *Fabricating the Keynesian Revolution: Studies of the Inter-war Literature on Money, the Cycle, and Unemployment*. Cambridge: Cambridge University Press.

_____. 2000. "Highlights of the Bullionist Controversy." Research Report No. 13. Stockholm: Institutet För Ekonomisk Historisk Forskning.

Myhrman, Johan. 1976. "Experiences of Flexible Exchange Rates in Earlier Periods: Theories, Evidence and a New View." *Scandinavian Journal of Economics* 78 (2): 169–96.

Niehans, Jurg. 1990. *A History of Economic Theory: Classic Contributions, 1720–1980*. Baltimore: Johns Hopkins University Press.

O'Brien, Denis P. 1975. *The Classical Economists*. New York: Oxford University Press.

Persson, Mats, and Claes-Henric Siven. 1993. "Pehr Niclas Christiernin." In *Swedish Economic Thought–Explorations and Advances*. Ed. L. Jonung. London: Routledge.

Ricardo, David. [1810–11] 1951. *The High Price of Gold Bullion. A Proof of the Depreciation of Bank Notes.* In *Works and Correspondence of David Ricardo*, vol. 3. Ed. P. Sraffa. Cambridge: Cambridge University Press.

_____. [1816] 1951. *Proposals for an Economical and Secure Currency*. In *Works and Correspondence of David Ricardo*, vol. 4. Ed. P. Sraffa. Cambridge: Cambridge University Press.

———. [1821] 1951. Letter to Wheatley. In *Works and Correspondence of David Ricardo*, vol. 9. Ed. P. Sraffa. Cambridge: Cambridge University Press.

Robbins, Lionel C. 1958. *Robert Torrens and the Evolution of Classical Economics*. London: Macmillan.

Salerno, Joseph T. 1980. The Doctrinal Antecedents of the Monetary Approach to the Balance of Payments. Ph.D. diss., Rutgers University.

Thornton, Henry. [1802] 1939. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. Ed. F. A. von Hayek. London: George Allen & Unwin.

Torrens, Robert. 1844. Letter II of *The Budget*: *A Series of Letters on Financial, Commercial, and Colonial Policy*. London: Smith, Elder and Co.

Viner, Jacob. 1937. *Studies in the Theory of International Trade*. New York: Harper Brothers.

Wicksell, Knut. 1918. "Inconveniences of the Measures for the Regulation of the Value of Money." *Svensk Handels-Tidning*, 25 September.

———. 1920. Memorandum on the High Prices. In *Report on the Question of How and to What Extent it is Possible to Set Up a Program for the Near Future for the Swedish Financial Policy*. Stockholm: Marcus: 39–63.

Accommodating Rising Population in Rural Areas: The Case of Loudoun County, Virginia

Raymond E. Owens and Pierre-Daniel G. Sarte

Washington, D.C., and Richmond, Virginia, are cities that share rich pasts in histories and politics. And although their centers lie 100 miles apart, the two areas also share something else—an approximately 12-mile-long border. According to the 2000 U.S. Census, the Washington, D.C., metropolitan statistical area (MSA) and the Richmond MSA literally bump into one another.

Although no one is likely to mistake the shared boundary area of the two MSAs for either city's downtown, MSAs have come to be the standard measure of a city's reach. That the two cities defined in this manner stretch well over 100 miles demonstrates the magnitude of population growth that has occurred in both urban areas. This growth is all the more impressive when one considers that much of it occurred in the last 30 years.

The rapid growth of suburban areas around Washington, D.C., and Richmond, Virginia (and of many cities like them), poses substantial challenges for both local elected officials and residents of those areas. These challenges include providing for housing, roads, sewers, schools, and the myriad requirements of a population spilling into formerly rural areas. Furthermore,

We wish to thank Yongsung Chang, Tom Humphrey, Daniel Tatar, and John Weinberg for many helpful comments on an earlier draft. We also thank Ben Mays, Clark Draper, Sean LaCroix, and Cindy Richmond for providing us with clarity on the issues facing and policies of Loudoun County. In addition, we greatly appreciate the research assistance provided by Elliot Martin and Matt Harris. Any errors are our own. The views expressed in this article do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

because the Commonwealth of Virginia provides limited guidance to localities concerning growth, these questions often must be addressed by local officials.

The inability of localities to address growth in a more aggressive manner is guided by the so-called Dillon rule.[1] The Dillon rule is a legal principle—used in Virginia—that addresses whether certain powers lie with local governments. The rule possesses two features. The first states that local governments have three types of powers: in layman's terms, those granted expressly by the state, those strongly implied by the state, and those that are essential to localities. The second part of the Dillon rule states that if there is any reasonable doubt whether a power has been conferred on a local government, then the power has not been conferred. This second feature effectively limits the fiscal tools available to localities to those strictly allowed by the state.

Attempts by officials of some counties to gain additional fiscal powers to fund the infrastructure required for an increasing population have had limited success at the statehouse in Richmond. As a result, the "toolkit" available to localities is often lacking in mechanisms that could prove useful in designing efficient growth policies. Perhaps because their tools are limited by law, localities have had to rely on available approaches such as taxes on real property, zoning, and cash proffers from residential and commercial developers—policies they *can* utilize—to stem the pressures from a rising population.

Although property taxes and zoning are generally well understood policies, proffers are lesser known. In short, proffers are payments made by developers to local governments as a part of a zoning or rezoning process. State law dictates that proffers are voluntary. The payments of the proffers may assist in gaining local government approval of the zoning action, but the law is clear that a zoning decision cannot be denied solely because a developer refused to pay a specified proffer amount. State law also specifies that proffers are not impact fees, though in practice they effectively approximate the latter. That said, development already zoned without proffers cannot legally be required to offset any impacts and, even in zoning cases where proffers are involved, the amount may not correspond to impact costs.

Zoning and cash proffers policies are not always popular with residents and developers in suburban counties. For example, in Loudoun County, Virginia, a largely rural county west of Washington, D.C., local policies to address population growth have occasionally reached a fever pitch. At a 1999 public hearing on Loudoun's growth policies held at the county courthouse, newspaper accounts described a near riot, noting that police officers had to be brought in to control the crowd. The episode prompted Thomas Sowell, a noted economic columnist, to devote a column to the issues facing Loudoun that ran in newspapers across the nation. But the vigorous debate over increasing

[1] For a history and more detailed description of the Dillon rule, see Writ (1989).

population in the county has also been closely watched by a number of groups interested in growth issues and by surrounding counties, all of whom view Loudoun's debate as a guide to the likely direction of policy in general.

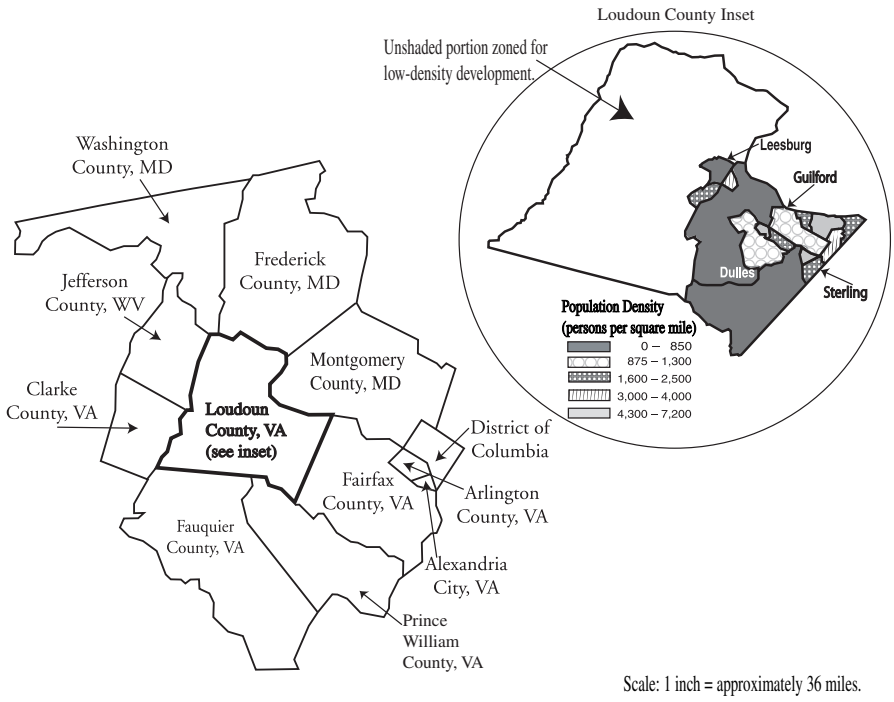
The Loudoun debate over population growth and how to address it is not surprising. Between 1990 and 2002, the county's population grew at a 7.5 percent annual rate, the second highest in the nation. But the rapid increase in the number of residents has not been welcomed by many in the county. County officials contend that the costs of infrastructure required to serve the population inflow exceeds the revenues generated, thus threatening the county's fiscal soundness (Meeting with Loudoun County Officials). In addition, incumbent residents complain that growth leads to more congestion. Yet there is less agreement as to the appropriate local policies to address the problems. Given the limited alternatives available, Loudoun officials have primarily adopted a zoning approach. Specifically, the county's board of supervisors has zoned the easternmost one-third of the county nearest Washington, D.C., (with the greatest population density) as residential, and it is available to accommodate additional population growth. In contrast, the westernmost two-thirds of the county (farthest from the downtown D.C. area) has been zoned for low-density development only, with allowable densities ranging from one household per 10 acres to one household per 50 acres (see Figure 1). These densities are so low that they effectively "shut the door" to new residential development in these areas of the county.

The situation in Loudoun raises many interesting questions concerning the effects of population growth on localities in the United States and in other industrial counties. Among the questions are, what kinds of impacts does population growth generate on county residents' welfare? Of the policies available to assist with the cost of population growth, which are used and why? And perhaps most important, are commonly used policies efficient in an economic sense?

The debate over these questions in Loudoun County, as in the broader debate, has been hampered by the lack of a formal model that identifies the likely relevant factors and traces out how they simultaneously affect residents' welfare. As a result, discussions between residents and local officials often isolate different arguments in an ad hoc manner, without providing a single coherent framework. The intent of this paper is to advance the debate by proposing a more formal treatment.

An examination of Loudoun's policies within a more formal setting may serve as a useful benchmark in analyzing the impact of rising population on counties. In particular, we will consider three responses to a rising population: zoning, raising the tax rate on real property, and using unrestricted proffers. We will discuss why these policies arise and explore whether they are efficient.

Figure 1 Population Densities in Loudoun County



To provide a framework in which these issues can be more systematically examined, this article presents a simple model of county agglomeration, inspired by Henderson (1987), where increases in population density lead to local congestion and higher prices for housing services. The model recognizes that opening a new area of a county to development entails substantial fixed costs linked to infrastructure construction and maintenance, such as sewage and water systems, highways, and schools, that are financed almost exclusively by local property taxes. Costs are fixed in the sense that opening an area to development requires a fixed amount of resources that is independent of the degree of residential development that takes place. Thus, an infrastructure network must typically be in place when an area is opened, irrespective of how many households actually move in. Under these conditions, we argue that localities' desire to maintain fiscal soundness combined with state legislated restrictions on their ability to raise revenues leaves them with little recourse outside zoning restrictions.

Since housing prices in a given region generally rise with population density, all else equal, opening new areas to residential development lowers the

average price of housing services as population spreads out across a larger area. The model then suggests that consumption of housing services generally rises but that the overall share of income devoted to housing remains unchanged, a result verified by data from Loudoun County. Hence, without sufficiently strong population growth, revenues from property taxes will fail to cover the additional cost of infrastructure associated with new residential developments. Indeed, Milligan (2003) argues that "one reason the Loudoun board used the blunt instrument of rezoning is because state lawmakers have resolutely refused to give localities other tools to manage growth." In other words, without the ability to acquire additional funds by raising property tax rates, the fixed costs associated with infrastructure construction and maintenance naturally lead to inertia in the creation of new residential developments.[2]

Because our model contains both property taxes and congestion externalities, the decentralized equilibrium is potentially inefficient. Even so, we show that in so far as congestion externalities are mainly local in nature, the decentralized distribution of individuals across locations is socially optimal. The presence of property taxes, however, does distort the consumption of housing services relative to other types of consumption. We argue that the policy of charging a proffer per housing unit to developers, which some localities have effectively followed in Virginia, constitutes a less distortional means of financing the costs of public infrastructure. In our framework, the use of proffers can actually help implement the first best solution in a decentralized setting.

1. A MODEL OF COUNTY AGGLOMERATION

Consider a county that encompasses $S > 0$ areas, where $\mathcal{S} = \{1, 2, \ldots S\}$, can be thought of as a group of Census tracts. We let $\mathcal{M} = \{1, 2, \ldots M\} \subseteq \mathcal{S}$ denote the set of areas open to residential housing. To be equipped for residential settlement, a region $i \in \mathcal{M}$ with land area $A_i > 0$ requires that a complete infrastructure network be provided and maintained. Examples of infrastructure include roads, sewer and water systems, schools, and public transportation, which in aggregate is assumed to carry a fixed resource cost, $\Phi(A_i)$, with $\Phi(0) = 0$, and $\Phi'(A_i) > 0$.

For now, we assume a fixed county population N, with N_i individuals living in location $i \in \mathcal{M}$. Each individual is endowed with one unit of labor, which he provides inelastically in a core city located outside the county. The distance from any area $i \in \mathcal{S}$ to the city varies depending on its location within the county. In Loudoun, for example, the relatively large area of the

[2] Although Loudoun has the legal authority to assess taxes against real and personal property and to accept proffers on housing created by new rezoning actions, officials stress that pressure from residents of the county and from state-level legislators limit their ability to raise these taxes to levels that would cover the cost and operation of infrastructure. Furthermore, the addition of new debt by the county is constrained by any deterioration in the county's revenue-expense ratio. The county finance director states that a less favorable ratio reduces the county's debt rating.

county means that some residents live as close as 25 miles from the center of Washington, D.C., while other residents may live as far away as 50 miles.

Production

Individuals are employed in the production of a county-wide traded good summarized by,

$$y = \lambda \sum_{i=1}^{M} N_i, \ \lambda > 0, \tag{1}$$

where y denotes the quantity of traded good output. These goods are produced competitively by firms that operate in the city center. Profit maximization by these firms immediately implies that

$$w_i = \lambda = w \ \forall i, \tag{2}$$

where w_i is the wage paid to individuals living in area i. Since individuals living in different regions of the county are perfect substitutes in production, they all earn the same wage. In the model below, the distribution of individuals across county areas derives from a tradeoff between commuting costs and the cost of housing services. To the degree that different individuals have different incomes, this tradeoff would involve net commuting costs instead. In a setting with net commuting costs, however, the substance of our analysis would remain largely unchanged.

Individuals living in different areas of the county open to development also consume housing which we treat as a location-specific good. As in Chatterjee and Carlino (2001), this good is produced using a technology that is linear in the traded good. Specifically, we have that

$$G_i = \left(\gamma d_i^{\eta}\right)^{-1} x_i, \ \gamma > 0, \ \eta > 0, \tag{3}$$

where x_i represents the quantity of the traded good required to produce G_i units of the local good in region i. The variable d_i denotes population density in location i, N_i/A_i. Thus, the factor γd_i^{η} in equation (3) captures the notion that higher population density reduces the efficiency of local good production.

Local good producers, interpreted here as providers of housing services, operate in a competitive market and maximize profits,

$$\max \ p_i G_i - x_i, \tag{4}$$

where p_i is the price of the local good in region i in units of the traded good. These producers take population density in each county area as given. Substituting for x_i in equation (4), and maximizing with respect to G_i, the price of the local good in county area i will then reflect its marginal cost,

$$p_i = \gamma \left(\frac{N_i}{A_i}\right)^{\eta}. \tag{5}$$

Therefore, as population increases in location i, so does the price of housing services in that area.

Preferences

Individuals that live in county location i have linear preferences over an aggregate good, C_i, given by

$$C_i = [(1-\delta_i)g_i]^{\theta} c_i^{1-\theta}, \tag{6}$$

where $0 < \theta < 1$ and $0 < \delta_i < 1$ $\forall i$. In equation (6), c_i and g_i represent consumption of the traded and local good, respectively. Since g_i represents housing services, individuals consuming g_i can be thought of as renters. The parameter δ_i captures the reduction in utility imposed by commuting between home and work. We can think of this reduction in the following way. Suppose two identical houses differ only in how far they are located from the workplace at the city center. A resident living at the more distant house spends more time commuting and correspondingly spends less time at home, thus getting less satisfaction (i.e., a higher δ_i) from a given amount of housing services. It is worth noting, though, that distance from the city center is not the only source of differences in commuting times. In practice, the location of roads, bridges, mountains, and physical features generally affect commuting times. Thus, it is entirely possible to find locations nearer to the city center that actually have longer commuting times to the city core. As a general rule, however, we expect that differences in housing services consumption will reflect the distance from the city center and the associated commute costs.

Each individual living in location i faces the following budget constraint,

$$p_i g_i + c_i \leq w - \tau p_i g_i, \tag{7}$$

where τ is a county-wide property tax that helps cover the cost of public infrastructure in areas open to development. In Loudoun County, the total real property and personal taxes collected each year amounts to $400 million, the approximate cost of operating the county's school system. Utility maximization subject to constraint (7) implies that a mobile individual residing in location i chooses

$$g_i = \frac{\theta w}{(1+\tau)p_i} \tag{8}$$

and

$$c_i = (1-\theta)w. \tag{9}$$

Equilibrium

We focus on equilibria where the distribution of individuals across county locations leaves no region open to development unoccupied. From equations

(3) and (9), the indirect utility achieved by an individual living in location i is,

$$\mathcal{V}_i = \theta^\theta (1-\theta)^\theta (1-\delta_i)^\theta (1+\tau)^{-\theta} p_i^{-\theta} w, \tag{10}$$

or, substituting for p_i and w,

$$\mathcal{V}_i = \theta^\theta (1-\theta)^\theta (1-\delta_i)^\theta (1+\tau)^{-\theta} \gamma^{-\theta} \left(\frac{N_i}{A_i}\right)^{-\theta\eta} \lambda. \tag{11}$$

Equilibrium with free movement of individuals requires that utility be equalized across all locations, $\mathcal{V}_i = \overline{\mathcal{V}}\ \forall i \in \mathcal{M}$. If there were a pair of locations i and j with $\mathcal{V}_i > \mathcal{V}_j$, individuals would seek to move from j to i. This would raise congestion in i and lower it in j until $\mathcal{V}_i$ and $\mathcal{V}_j$ were equalized. In addition, the sum of individuals across locations open to development must equal the exogenous county-wide population,

$$\sum_{i=1}^{M} N_i = N. \tag{12}$$

Finally, the county must cover the fixed costs associated with providing and maintaining public infrastructure in the developed areas,

$$\sum_{i=1}^{M} N_i \tau p_i g_i = \sum_{i=1}^{M} \Phi(A_i), \tag{13}$$

where the left-hand side of the above expression denotes tax revenues from property taxes.

Proposition:

Under the maintained hypotheses, there exists a unique distribution of individuals across open locations, $N_i,\ i = 1,\ldots,M$, with common utility, $\overline{\mathcal{V}} > 0$.

Proof:

Observe that the conditions $\mathcal{V}_i = \overline{\mathcal{V}}\ \forall i \in \mathcal{M}$ and $\sum_{i=1}^{M} N_i = N$ make up $M+1$ equations in $M+1$ unknowns, namely $N_i,\ i = 1,\ldots,M$, and $\overline{\mathcal{V}}$. Thus, rewrite equation (11) as

$$N_i = \left[\frac{\overline{\mathcal{V}}}{\theta^\theta (1-\theta)^\theta (1-\delta_i)^\theta (1+\tau)^{-\theta} \gamma^{-\theta} \lambda}\right]^{-\frac{1}{\theta\eta}} A_i.$$

Substituting this expression into equation (12), it follows that $\overline{\mathcal{V}}$ must solve

$$\sum_{i=1}^{M} \left[\frac{\overline{\mathcal{V}}}{\theta^\theta (1-\theta)^\theta (1-\delta_i)^\theta (1+\tau)^{-\theta} \gamma^{-\theta} \lambda}\right]^{-\frac{1}{\theta\eta}} A_i = N.$$

Define the left-hand side of the above expression as $F(\overline{\mathcal{V}})$, and note that $\lim_{\overline{\mathcal{V}}\to 0} F(\overline{\mathcal{V}}) = \infty$ while $\lim_{\overline{\mathcal{V}}\to\infty} F(\overline{\mathcal{V}}) = 0$. Since $F(\overline{\mathcal{V}})$ is continuous, by

the Intermediate Value Theorem, there exists $\overline{\mathcal{V}} > 0$ such that $F(\overline{\mathcal{V}}) = N$. In addition, because $F(\overline{\mathcal{V}})$ is strictly decreasing in $\overline{\mathcal{V}}$ on $[0, \infty)$, this solution is unique.

Given the solution for $\overline{\mathcal{V}}$, one can then simply solve for the distribution of individuals across location using (11).

The model of county agglomeration we have just presented possesses two important features that emerge as equilibrium outcomes.

First, the relative price of housing services between any two county areas reflects differences in commuting costs. In particular, from equation (10), the condition that $\mathcal{V}_i = \mathcal{V}_j$ for any two areas open to residential housing implies that

$$p_i = p_j \left(\frac{1 - \delta_i}{1 - \delta_j} \right) \quad \forall i \text{ and } j \in \mathcal{M}. \tag{14}$$

In other words, in choosing where to live within the county, individuals will trade off the price of housing services against commuting costs. In particular, county locations that involve a shorter commute to work will tend to have higher-priced housing services. In fact, this result appears to hold in Loudoun, though the heterogeneity of the housing stock makes a precise measure difficult. According to county officials, identical houses in areas with lower commuting costs generally command higher prices (and thus rents) than similar houses in areas of the county with higher commuting costs.

Second, because prices of housing services reflect congestion externalities driven by higher density, county areas with higher commuting costs will also have lower densities. In (14), $\delta_i < \delta_j$ implies that $p_i > p_j$. By equation (5), we then also have that $d_i > d_j$.

At this stage, we find it useful to introduce a numerical example to better highlight key features of our model as the economic environment changes. Specifically, given the debate surrounding Loudoun County, we focus on the effects of a rising county population as well as those of a change in the number of areas open to residential housing. We shall also use this numerical example below in making comparisons with the efficient solution.

Calibration to Current Loudoun County Benchmarks

According to our model, differences in commuting costs, δ_i, lead to varying densities in different regions. Therefore, as shown in Figure 1, we partition the developed eastern region of Loudoun County (i.e., the region unaffected by zoning restrictions) into density quintiles and set δ_i to match the density of each of the five areas. The associated five land areas have sizes, in square miles, 116.4, 27.3, 13.2, 8.2, and 5.7, and we calibrate A_i to match each of these land areas. The population density in these five areas are, in people per square mile, 208.21, 1,072.71, 1,909.00, 3,563.51, and 5,613.52. Observe in Figure 1 that low-density areas tend to be farther away from Fairfax County

Table 1 Model Parameters

	Calibrated Benchmark Parameters	Value
Preferences		
θ	Housing share of income	0.25%
Technology		
λ	Per capita income	\$50,238
γ	Scalar in density congestion	93.41
η	Curvature in density congestion	0.49
Geography		
δ_i	Commuting costs	[0.09, 0.11, 0.14, 0.17, 0.23]
A_i	Land area (square miles)	[5.7, 8.2, 13.2, 27.3, 116.4]
N	Population	139,873

and Washington, D.C., where Loudoun County residents typically commute to work.

According to the U.S. Census, the population in the developed areas of Loudoun County currently stands at 139,873, and we set N to match this value. We choose λ to reflect an individual's yearly earnings in the county, \$50,238. This number reflects a weighted average of male and female full-time workers. From the Census, the share of income spent on housing and property taxes, θ, is approximately 0.25. It is difficult to get an accurate housing price per square foot corresponding to each region, where we think of square footage as a proxy for housing services. However, data from the Loudoun County Office of Mapping and Geographic Information suggests that \$143 per square foot is a reasonable upper bound for that county. We then choose γ to match this upper bound in equilibrium, $\gamma = 93.41$, and set η assuming a 15 percent gradient in housing prices from the most to the least dense area.

Finally, because individuals spend $p_i g_i$ of their yearly disposable income on housing services, current housing values, V, for the typical individual are given by

$$\begin{aligned} V &= \phi\,(p_i g_i)\,, \\ \phi &= \frac{1}{r}\left[\frac{(1+r)^{T+1}-1}{(1+r)^{T}}\right], \end{aligned} \tag{15}$$

where ϕ is a factor that captures the present value of a one dollar annuity discounted over the number of years that a house provides services, T, and rate, r. In particular, given that the typical household contains 2.7 individuals in Loudoun County, our model suggests that the representative house is worth approximately \$435,000 when $T = 30$ and $r = 0.05$. Property tax rates in Loudoun County are currently set to 1.08 percent of housing values. Since

Table 2 Model and Data Statistics

Population Distribution	
Loudoun County	[24,227; 29,265; 25,226; 29,348; 31,807]
Model	[24,231; 29,238; 25,235; 29,375; 31,791]
Density Distribution	
Loudoun County	[208.21; 1,072.71; 1,909.35; 3,563.51; 5,613.52]
Model	[208.24; 1,071.79; 1,910.36; 3,569.32; 5,616.88]
Average Housing Prices	
Loudoun County	$427,199
Model	$435,279
Household Property Taxes	
Loudoun County	$4,613.00
Model	$4,701.00

the property tax in (7) applies to $p_i g_i$ rather than $\phi\,(np_i g_i)$, where n is the number of individuals per household, we let $\tau = 0.0108\phi n$, or 0.18. This tax generates about $4,700 per household yearly. The parameters that achieve our calibration targets are summarized in Table 1.

Table 2 reports the model-generated population and density distribution in each of the five areas depicted in Figure 1. As shown in the table, the model, although stylized, does well in reproducing actual Loudoun County statistics. In addition, we are also able to approximate statistics we had not explicitly targeted. For instance, both average housing prices and yearly property taxes collected per household conform relatively well to the data.

Zoning Restrictions in the Face of Increasing Population

The model above implies that in a given year, approximately $243 million are collected in property taxes in the developed region of Loudoun County.[3] Since this revenue is used exclusively to finance the provision and maintenance of public infrastructure, the corresponding fixed costs come to slightly more than $1.42 million per square mile.

[3] According to the 2000 Census, in total, Loudoun County collects $300 million in real property taxes, approximately $223 million of which, close to the model's prediction, comes from the eastern, developed portion of the county.

Suppose that local authorities were to consider lifting zoning restrictions on 10 additional square miles adjacent to the already developed part of the county. Because we assume this area to be immediately adjacent to the least dense populated region, we posit similar commuting costs, $\delta = 0.23$. Residents of the new area, therefore, would incur commuting costs equivalent to a 23 percent reduction in housing services. Moreover, according to our calculations, opening this region to development would require an additional $14.2 million in property taxes.

With no population growth and no adjustment in property tax rates, our model implies that total property taxes collected would also remain unchanged. The existing population would spread out across a larger area thus lowering density and, by equation (5), local goods' prices. However, by equation (8), individuals would then increase their consumption of housing services so as to leave the share of income they spend on housing exactly unchanged. In practice, the share of income devoted to housing services is indeed nearly constant over time, not only in Loudoun County and Virginia, but nationally. Since this amount helps determine housing values in equation (15), it follows that these values would then remain unaffected and so would the resulting property taxes. Hence, opening a new area to residential housing is feasible only if the rate of net migration into the county is sufficient to generate tax revenues equal to the additional fixed costs incurred.

In our hypothetical example, the existing county population would have to increase by approximately 5.9 percent to yield an increase in the tax base large enough to generate an additional $14.2 million. This represents an increase of around 8,250 individuals or 3,050 households. Thus, given that installing and maintaining infrastructure entails substantial fixed costs, our analysis implies inertia in the creation of new developments. That is, the population residing in areas already open to development has to reach a high enough threshold that the tax base can cover the additional cost of new infrastructure. Therefore, with legislated and/or political limits on a county's ability to raise property tax rates, local authorities have little practical recourse other than to appeal to low-density zoning restrictions. Note that while population grows to meet a threshold that would allow the county to open a new area, density increases, local goods prices rise, and consumption of housing services fall. Consumption of the aggregate good, C_i, in equation (6), therefore, decreases for the representative individual.[4] It is no surprise, therefore, that some county residents complain of congestion and exert pressure on Loudoun County's board of supervisors to lift zoning restrictions.

It is important to note that given a fixed county population, N, opening a new area of the county to residential development in our framework does not necessarily increase welfare. On the one hand, the new area would allow for lower population density and lower congestion in existing regions of the

[4] Recall that all individuals have the same utility in equilibrium, $C_i = C_j \ \forall i, j \in \mathcal{M}$.

county. This effect induces individuals to consume more housing services which increases welfare. On the other hand, to the degree that the cost of additional infrastructure raises property tax rates, consumption of housing services would fall. On net, it is not clear that consumption of housing services would increase if an additional region of the county were zoned for residential settlement. Furthermore, from (6), commuting costs associated with the new area would also play a direct role in the evaluation of welfare.

In Loudoun County, rates of increase in the county's population are forecast to remain high, though not as high as in the 1990s. Loudoun's Department of Economic Development projects that the county's population will rise at an average annual rate of 4.7 percent over the next nine years, about triple the rate of population growth expected in the United States over the same period. Thus, it is likely that an imbalance between population growth, revenues, and infrastructure adequacy will continue to face the county.

2. THE SOCIAL PLANNER'S PROBLEM

We now show that the outcomes in the decentralized county economy are not Pareto optimal. Specifically, we can assess Pareto optimality by comparing our results above with the results from the same problem for a hypothetical social planner. Contrary to most models in regional economics, the source of inefficiency in our framework does not stem from the congestion externalities linked to density. In our model, these externalities are local in nature and, therefore, directly reflected in the price of housing services in the concerned region. In essence, the technology in (3) assumes that greater density in region i congests the production of housing services in that region, and not in another region that is further away.[5] The population density distribution, therefore, replicates that which emerges in the decentralized equilibrium. The presence of county taxes, however, does distort the consumption of housing services relative to other types of consumption. We argue that local government should finance infrastructure by charging developers a lump sum proffer per housing unit rather than relying on property taxes. Some localities in Virginia charge proffers. We show that their approach can actually help implement the first best solution in the decentralized setting.

The social planner looks to maximize the utility of households in the county, as given by

$$\sum_{i=1}^{M} N_i[(1-\delta_i)g_i]^{\theta} c_i^{1-\theta}. \tag{16}$$

[5] This is also the case in Chatterjee and Carlino (2001).

The only constraints faced by the planner are the county's resource constraint,

$$\sum_{i=1}^{M} N_i c_i + \sum_{i=1}^{M} N_i x_i + \sum_{i=1}^{M} \Phi(A_i) = \lambda \sum_{i=1}^{M} N_i, \tag{17}$$

and the requirement that population in regions open to development add up to county population, (12). The middle term on the left-hand side of (17) captures the resource costs, in units of the traded good, associated with the county-wide provision of housing services, where x_i is implicitly defined by the technology in (3).

The planner's optimal choice of regional traded good consumption, c_i, local good consumption, g_i, and regional population, N_i, are respectively given by

$$(1-\theta)[(1-\delta_i)g_i]^{\theta} c_i^{-\theta} = \mu_2, \tag{18}$$

$$\theta[(1-\delta_i)g_i]^{\theta-1}(1-\delta_i)c_i^{1-\theta} = \mu_2 \gamma \left(\frac{N_i}{A_i}\right)^{\eta}, \tag{19}$$

and

$$[(1-\delta_i)g_i]^{\theta} c_i^{1-\theta} + \mu_2 \left[\lambda - c_i - \gamma(1+\eta)\left(\frac{N_i}{A_i}\right)^{\eta} g_i\right] - \mu_1 = 0, \tag{20}$$

where $\mu_1 \geq 0$ and $\mu_2 \geq 0$ are the Lagrange multipliers associated with constraints (12) and (17).

We now demonstrate that the planner's solution entails the same distribution of population across regions as that found in the decentralized equilibrium. To see this, observe first from (14) that the decentralized allocation of individuals across regions can be summarized by

$$(1-\delta_j)\gamma^{-1}\left(\frac{N_j}{A_j}\right)^{-\eta} = (1-\delta_i)\gamma^{-1}\left(\frac{N_i}{A_i}\right)^{-\eta} \quad \forall i \text{ and } j \in \mathcal{M}. \tag{21}$$

Under the optimal solution, we can use equations (18) and (19) to show that

$$(1-\theta)\theta(1-\delta_i)\gamma^{-1}\left(\frac{N_i}{A_i}\right)^{-\eta} = \mu_2^{\frac{1}{\theta}} \quad \forall i \in \mathcal{M}. \tag{22}$$

Since μ_2 is constant across regions, equation (22) implies (21), and the optimal allocation of individuals across locations replicates that of the decentralized equilibrium. Because in our model, congestion externalities reduce the production efficiency of housing services locally, individuals who move and congest a given region have to pay higher prices for housing services in that region. As in Chatterjee and Carlino (2001), the formulation of local externalities seems to us more reasonable than one where a region's density

decreases the production efficiency of housing services in another area that is potentially much further away.[6]

It remains that in the decentralized equilibrium, the presence of taxes on housing services distorts the allocation of consumption between the traded and local good. In our model, this distortion is small and results only in a 0.2 percent loss in welfare when measured in terms of the aggregate consumption basket, C_i. However, we now argue that allowing localities to charge developers a lump sum proffer to finance public infrastructure can help remove the distortion altogether.

Using Lump Sum Proffers as a Means to Finance Infrastructure

The main trouble with county taxes, as depicted in (7), is that they are proportional to housing services—and thus housing values—which leads to suboptimal decentralized allocations. In other words, individuals in every locality are led to consume less housing services than they otherwise would absent taxes. Historically, however, localities in Virginia have had the ability to accept voluntary lump sum cash proffers from residential developers, independent of the quantity of housing services they provide.[7] The courts in Virginia have held that the absence of "voluntary" payments cannot be the sole reason for denying zoning or rezoning. However, many counties, including Loudoun, publicize the recommended proffers per residential housing unit constructed. In a setting where a new area is opened to development, all houses constructed would be subject to a lump sum proffer. In the case of opening a new area to housing, zoning, or rezoning action would be necessary so that proffers could apply to all housing.

We now show that, in the decentralized equilibrium, these proffers would simply be passed on to consumers, provided developers operate in a competitive market. More importantly, because they are non-distortionary, using these proffers in lieu of property taxes would allow the market equilibrium to replicate the social optimum.

Suppose that each county locality charges developers a cash proffer, Π_i, per housing unit that is unrelated to the amount of housing services they sell, G_i.[8] In equilibrium, these cash proffers have to be such that $\sum_{i=1}^{M} N_i \Pi_i = \sum_{i=1}^{M} \Phi(A_i)$ to maintain the feasibility of areas open to development. Let

[6] Note that the social optimum would yield population allocations different than those given by the decentralized equilibrium if congestion had an effect on commuting costs.

[7] In practice, these proffers can only be raised following zoning or rezoning actions.

[8] Observe that housing units can be of different sizes, our proxy for G_i. Furthermore, an implicit assumption here is that each individual requires one housing unit, although individuals can combine into households.

$R(G_i) = p_i G_i + \Lambda_i$ denote a developer's revenue from selling G_i units of housing services in locality i. Developers are assumed to operate in a competitive market, and we allow for any pricing rule that enables firms to charge both a price per unit of housing services, p_i, and a fixed amount, Λ_i, that could potentially be zero.[9]

From (3), a developer's profits in terms of the traded good are given by

$$R(G_i) - \gamma d_i^{\eta} G_i - \Pi_i. \tag{23}$$

It is then easy to see that the pricing rule whereby firms charge γd_i^{η} per unit of housing services and pass on the entire cash proffer to consumers constitutes a unique equilibrium pricing rule. First, to see why it is an equilibrium rule, observe that a firm with a pricing strategy such that $R(G_i) > \gamma d_i^{\eta} G_i + \Pi_i$ would have no customers. Other firms would be able to charge slightly less and capture the entire demand while still making at least zero profits. On the other hand, a pricing rule that yielded revenues less than $\gamma d_i^{\eta} G_i + \Pi_i$ would have the firm make negative profits and is not sustainable. Therefore, in equilibrium, firm revenues have to be exactly $\gamma d_i^{\eta} + \Pi_i$. Second, to see why $\{p_i, \Lambda_i\} = \{\gamma d_i^{\eta}, \Pi_i\}\ \forall i \in \mathcal{M}$ represents a *unique* equilibrium pricing rule, consider any other strategy, $\widetilde{p}_i = \gamma d_i^{\eta} + \varepsilon$, $\varepsilon \gtrless 0$ and $\widetilde{\Lambda}_i$. Because total revenue must be $\gamma d_i^{\eta} + \Pi_i$ in equilibrium, a firm that charges $\widetilde{p}_i$ per unit of housing services would have to adjust the fixed portion of its pricing strategy such that $\widetilde{\Lambda}_i = \Pi_i - \varepsilon G_i$. But this contradicts the notion that $\widetilde{\Lambda}_i$ is independent of G_i. Therefore, the rule whereby firms charge marginal cost per unit of housing services and pass on the entire cash proffer required by the county to individuals is the only equilibrium pricing rule.[10]

Of course, the main point here is that faced with this pricing rule, individuals' budget constraint (7) in the decentralized county economy becomes

$$(p_i g_i + \Lambda_i) + c_i \leq w. \tag{24}$$

Hence, their consumption of housing services is no longer distorted relative to other types of consumption, and the decentralized equilibrium can achieve the first best solution. Observe that while individuals pay more for housing services relative to the previous section, they no longer have to pay taxes on housing services. In fact, since all that matters in terms of providing county-wide infrastructure and its operation is that its costs be covered by cash proffers collected from developers, $\sum_{i=1}^{M} \Phi(A_i) = \sum_{i=1}^{M} N_i \Pi_i$, the county can design a regional distribution of proffers such that the difference between what individuals now pay for housing services and what they paid in the

[9] Since our model is static and does not distinguish between housing stock built at different dates, we think of Λ_i as the yearly amount corresponding to the capitalized proffer value that a developer would be charged at the time of construction.

[10] Observe also that any two-part pricing strategy that successfully attracts customers away from $\{p_i, \Lambda_i\} = \{\gamma d_i^{\eta}, \Pi_i\}$ necessarily yields negative profits.

previous section exactly equals what they were originally spending in property taxes, $\Lambda_i = \Pi_i = \tau p_i g_i$. There is no sense, therefore, in which this proffer-based policy would ultimately end up being more costly to individuals.

If proffers allow counties to offset the cost of infrastructure and its operation associated with new housing, why is zoning still used along with proffers in Loudoun County? The answer lies in the legal restrictions associated with proffers. Legally, proffers can be used to offset fully or partially only the capital costs of infrastructure, not operating costs. In the case of schools, for example, the operating cost is a substantial portion of the total cost, meaning that proffers will not overcome the fixed cost problem discussed earlier. Without the ability to use proffers to offset infrastructure costs fully, counties resort to zoning to limit the fiscal impact of rising population.

3. SUMMARY REMARKS

Rapidly increasing population in formerly rural counties on the fringe of urban areas has strained local governments' ability to provide infrastructure, raising congestion levels. The difficulty in providing adequate infrastructure lies in the fixed cost nature of infrastructure production as well as political and legal restrictions on localities' ability to raise revenue. Using a simple model of locational choice, we find that local officials could best balance population and infrastructure through a lump-sum proffer fee on developers. Provided that the market in which developers operate is competitive, this impact fee is likely to be passed onto users of housing services. This approach has the well-known advantage of being non-distortionary with respect to individuals' consumption decisions. Alternatively, balancing infrastructure and population can be achieved through setting an appropriate real property tax, though this approach introduces distortions into individuals' consumption decisions, leaving them with less aggregate consumption than with lump sum fees.

In addition, we find that legal and political restrictions on county officials' use of proffers and real property taxes have led them to the use of zoning in practice. Given the substantial fixed costs associated with infrastructure provision and the use of zoning, rising population leads to increased congestion before the number of households reaches a high enough threshold to make it feasible to open up a new land area. Ultimately, however, zoning remains an inefficient means to address localities' infrastructure and population issues. A more efficient solution would be to lessen restrictions on localities' use of proffers and their ability to raise revenue more generally.

Although Loudoun County, Virginia, has been used as a case study for calibration of our model, the framework set out in this paper should be broadly applicable to the problem associated with rising population in many areas of the United States. Indeed, the fixed cost aspect of infrastructure provision

combined with restrictions on localities' ability to raise revenue appear to be applicable to localities broadly.

REFERENCES

Chatterjee, S., and G. Carlino. 2001. "Aggregate Metropolitan Employment Growth and the Deconcentration of Metropolitan Employment." *Journal of Monetary Economics* 48 (December): 549–83.

Henderson, V. 1987. "General Equilibrium Modeling of Cities." In *Handbook of Regional and Urban Economics,* vol. II*: Urban Economics*. Ed. E.S. Mills. New York: New Holland.

Loudoun County Office of Mapping and Geographic Information. http://www.loudoun.gov/omagi/index.htm (accessed December 23, 2003).

Meeting with Loudoun County Officials (Ben Mays of Loudoun County Management Services and Clark Draper, Sean LaCroix, and Cindy Richmond of the Loudoun County Department of Economic Development), Leesburg, Virginia. August 2, 2003.

Milligan, J. 2003. "Showdown in Loudoun." *Virginia Business* (April): 8–13.

Sowell, T. 2001. "Property Rights." Townhall.com. http://www.townhall.com/columnists/thomassowell/ts20010809.shtml (accessed December 23, 2003).

U.S. Census Bureau. "United States Census 2000." http://www.census.gov/main/www/cen2000.html (accessed December 23, 2003).

Writ, Clay L. 1989. "Dillon's Rule." *Virginia Town and City* 24 (8): 12–15.

Closing Troubled Banks: How the Process Works

John R. Walter

Business failure typically occurs when a financially weak firm can no longer pay its creditors. Failure generally involves a series of steps. First, the firm suffers losses. Second, when the firm's creditors learn of the losses, they increase their estimate of the firm's probability of default. To compensate themselves for this increased risk, creditors demand higher interest rates or require debt repayment. Third, the firm finds itself unable to raise or generate additional funds to meet those demands and defaults. Creditors then either force the firm into bankruptcy, in which case a bankruptcy court decides how to best allocate the firm's assets to meet its debts, or the firm privately arranges with creditors for a payout of firm assets. In either case, the assets can be redeployed in more valuable uses. While business failure is often exceptionally disruptive for the firm's managers and employees, it is beneficial for society since it ensures that business resources are not devoted to ineffectual enterprises.

But what about banks? How does their failure ensue? A high proportion of bank liabilities are government-insured deposits. Deposit insurance primarily exists to prevent inappropriate bank runs that may occur when many of a bank's depositors seek to withdraw funds even though the bank is healthy. While it solves one problem—inappropriate runs—deposit insurance can create another. Insured depositors have no incentive to demand higher interest rates or debt repayment when their bank is troubled. While uninsured bank creditors will likely demand repayment, the bank can often raise funds to meet these demands by gathering new insured deposits at little extra cost. As a result, the market-driven process of failure and subsequent reallocation of assets is short-circuited for banks, and its societal benefits are muted.

Because market forces are unlikely to bring about the timely closure of troubled banks, the government agencies that charter and supervise banks

The author benefited greatly from discussions with Kartik Athreya, Marvin Goodfriend, Tom Humphrey, and John Weinberg. The views expressed herein are not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

are typically left to decide when a bank is no longer viable and should be closed. Mistakes by the agencies can create significant inefficiencies. For example, during the 1980s many insured depository institutions remained open long after they became insolvent. As a result, financial resources were tied up in inefficient operations for extended periods. Legislators recognized the problem and in 1991 enacted the Federal Deposit Insurance Corporation Improvement Act (FDICIA). The Act required bank supervisors to step in and close depository institutions more quickly and reformed the process by which the Federal Deposit Insurance Corporation disposed of failed depositories. All of which raise the crucial question: How do these agencies decide when to step in under rules established by the FDICIA? Further, how do the agencies proceed following intervention? This article seeks to answer these questions.

Supervision of healthy banks is intended to ensure that the government safety net does not provide incentives for banks to undertake inefficient risks. Supervision of banks includes propagating and enforcing restrictions on risky activities, setting deposit insurance premia and insurance coverage levels, and establishing minimum capital requirements as well as examining banks to ensure that capital requirements are met. The supervisory treatment of failing banks, however, is likewise important if the safety net is to be prevented from inducing excessively risky behaviors.

This article provides an overview of the process by which troubled banks are closed. It not only points out the beneficial changes to the FDICIA's closure process, but also indicates some areas of remaining weakness. One example involves the opportunities for uninsured depositors to escape losses by withdrawing their deposits immediately prior to bank failure. The FDICIA addresses one avenue of escape by restricting Federal Reserve discount window lending to troubled banks. As will be shown, even if the Act prevented all such lending, many uninsured depositors would be likely to escape losses nevertheless. Likewise, the oft-discussed systemic risk provisions of the FDICIA can provide incentives for excessive risk-taking.

Creditors of troubled nonbank firms have every incentive to require their prompt closure and to ensure that the assets of such firms are redirected. Typically bank supervisors must perform these roles for troubled banks. The FDICIA attempts to ensure that supervisors have incentives to close banks promptly and to dispose of their assets efficiently. This article is intended as an aid to those who would like to understand and perhaps improve the bank closure process.

1. NONBANK FAILURE VERSUS BANK FAILURE

In broad terms, failure occurs when a firm becomes unable to produce a market rate of return on its owners' capital investment. In other words, earnings from operations are insufficient to meet expenses, including interest payments to debtholders, and to pay owners a market return. The unproductive firm's assets

might simply be sold off, debts repaid, and the business closed. In such an example, losses are borne only by owners, with debtholders escaping any loss.

While closure can occur without great losses, frequently firm closure occurs under less favorable circumstances. Severe losses may rapidly overwhelm the firm. Alternatively, owners may not realize the severity of the earnings decline, or they may choose to hold out, hoping for a recovery. In such a situation losses can become large enough to consume a significant portion of the firm's assets. Owners are likely to seek additional funding from lenders or new investors. These providers will attempt to ascertain the likelihood that the firm can recover and produce a market return. If a recovery is deemed probable, perhaps including reorganization, then creditors advance new funds. Alternatively, if lenders or investors view the firm's troubles as long-lasting they will be unwilling to lend. The firm then is unable to meet payment demands from its creditors. These creditors can then force the closure of the firm in bankruptcy or by foreclosing on its assets.

In some cases, banks may follow the first discussed pattern of closure. Specifically, bank owners may find that their operation is no longer producing a market rate of return and that their investment can earn a higher return employed elsewhere. In such a case, the owners may simply liquidate the bank. Between 1991 and 2001, 38 banks were liquidated without the supervisors stepping in and requiring closure (FDIC 2003).[1] More typically, owners may sell their assets and liabilities to another bank. While bank supervisors do not classify such outcomes as failures, these outcomes are driven by the realization by bank owners that returns are insufficient. So liquidations and some mergers meet the definition of an economic failure.

Thus, while a bank failure can be similar to the failure of a nonbank firm so that market players—specifically the bank's owners—close the bank without supervisory intervention, in general a bank failure producing losses beyond those suffered by equity holders will be very different. The reason is that deposit insurance allows a bank with weak earnings to continue to raise funds sufficient to cover its losses. Imagine a bank that has suffered a string of losses, such that it would be clear to outsiders that the bank is unlikely to be able to repay out of earnings any additional debts it undertakes. A nonbank firm in such a situation would be unable to convince investors to extend funds to cover future operations. But depositors, protected against loss by a government guarantee of repayment, are quite willing to provide the bank with additional funding. As a result, insured-troubled banks can frequently remain open long past the time they would be closed absent deposit insurance.

[1] This figure includes banks that discontinued deposit operations or otherwise liquidated or closed. It excludes banks that failed or merged. See "Notes to Users" in FDIC (2003). While the bank supervisors did not force the closure of these banks, it is likely that the threat of supervisory intervention may have been important in some of the closures. Owners may have shunned closure had the threat not been looming.

This distinction between investors' treatment of nonbanks and depositors' treatment of banks points out why banks are special, and why bank supervisors must be especially vigilant to quickly close troubled banks. Once any firm, whether bank or nonbank, suffers losses great enough to consume all of its owners' investment, its owners have a strong incentive to undertake very risky investments, since at that point they have no more investment to lose. Risky investments will be attractive because they might produce a large enough return to save the firm. While investors will be unwilling to extend to nonbanks the funds needed for such risky investments, because of deposit insurance, banks can gather funding. Consequently, the failure to close a troubled bank quickly, affords bank owners and managers the ability and incentive to undertake risky, high-loss investments.

2. WHO LOSES WHEN BANKS FAIL?

When nonbank firms fail, creditors often are not repaid in full and can suffer large losses as a proportion of their investment. When banks fail, insured depositors, who provide the largest portion of the typical banks' liabilities, are protected against losses by the FDIC. Bank shareholders are not insured by the FDIC and will normally suffer losses, and uninsured depositors, to be discussed in more detail later, also may lose.

The FDIC promises to insure deposits up to $100,000 per account. Backing the promise is the FDIC's reserve fund, which, as of the end of 2002, summed to $32 billion, or 1.27 percent of all insured deposits in banks. The fund was accumulated and is maintained from insurance premiums charged banks and from the interest earned on the reserves. A separate reserve is maintained by the FDIC for savings associations. Relative to insured deposits, it is of a similar size to the bank fund and is also funded by premiums charged these associations. If the bank reserve fund falls below a specified minimum relative to insured deposits—1.25 percent of insured deposits—banks are required to rebuild the fund through higher premiums, as are savings associations.

But what if payments necessary to protect depositors grow so large that they deplete the entire fund? In this case the FDIC can draw on a $30 billion line of credit from the U.S. Treasury, to be repaid by future premiums. Beyond this amount, taxpayers would make up the difference. During the savings and loan crisis of the 1980s, the reserves backing savings association deposit insurance were depleted. In 1987, a statute was enacted promising Treasury backing, and therefore taxpayer backing, for insured deposits. Over the next several years taxpayers provided over $100 billion to protect depositors in savings associations from any losses.

3. WHO CLOSES BANKS?

So, if deposit insurance gives creditors little incentive to foreclose on a troubled bank or force it into bankruptcy, who does force its closure? The answer is

that a troubled bank is closed by the agency which granted the bank a charter to operate. The agency appoints an entity, typically the FDIC, to act as a receiver or conservator. This entity then takes control of the bank's assets and liabilities and ultimately divides the assets among liability holders. For national banks, charters are granted and the closure decision made by the Office of the Comptroller of the Currency—a bureau of the U.S. Department of the Treasury—which also supervises national banks. For state-chartered banks, charters are granted, and supervision is provided, including the decision to close troubled banks, by an agency of the state government. Similarly, some savings associations have a federal government-granted charter and are supervised and closed by the Office of Thrift Supervision, also a bureau of the Treasury; others have state charters.

In the case of national banks and federally chartered savings associations, federal law requires that the receiver be the FDIC (12 U.S. Code 1821c). Most states also require their banking agencies to name the FDIC as receiver (FDIC 1998b, 69).

While state banks are typically placed in receivership by a state banking agency, in some cases a federal bank supervisor can make the decision to appoint a receiver for a state-chartered bank. The Federal Reserve or the FDIC, neither of which grant bank charters, share with state banking agencies the duty of supervising state-chartered banks, including the authority, under federal law (discussed below), to appoint a receiver. The backup authority granted the federal agencies could be attributed to the fact that only the federal agencies answer to the federal legislature, which must appropriate the funds to protect depositors should the reserve funds be depleted. On occasion the FDIC has used its authority. For example, over the decade from 1993 through 2003, out of a total of 54 state-chartered bank failures, the FDIC closed three state-chartered banks without the concurrence of the state banking agency. Such closures can imply a disagreement between the state agency and the FDIC concerning the severity of the troubled bank's problems; the state agency concludes that the bank is viable, while the FDIC concludes that the bank is not.

4. CLOSURE POLICIES

Policy guiding regulators' approach to troubled banks is established by the *Prompt Corrective Action* (PCA) rules, found in Section 131 of the FDICIA. The rules primarily focus on the level of bank capital, meaning the dollar amount by which assets exceed liabilities, i.e., the net worth or solvency of the bank. According to the PCA rules, the federal bank regulatory agencies are required to appoint a receiver or conservator for the bank within 90 days of the time the bank's capital-to-assets ratio falls below two percent (12 U.S. Code 1831o). In other words, if a bank becomes insolvent, or close to it, a

receiver must be appointed. Alternatively, the federal agencies can choose another course, such as granting the bank additional time or injecting funds to prop up the troubled bank, but such a decision must be reviewed every 90 days until the bank has recovered. After 270 days, if the bank remains undercapitalized, a receiver must be appointed unless the bank has positive net worth and its financial health is clearly improving.

Other provisions of federal law provide additional grounds by which agencies may step in. For example, a banking agency may appoint a receiver even though the bank may not currently have weak capital if the agency determines that the bank is likely to incur losses that will deplete substantially all of its capital. Outside of capital considerations, a receiver may be appointed if the bank has willfully violated a cease and desist order (12 U.S. Code 1821c).

5. CLOSURE IN PRACTICE

But how do the state or federal regulators learn of declining health of banks, and what steps do they typically take on the way to determining that the bank is no longer viable? Two means are typically employed by banking agency examiners to measure the health of a bank, namely off-site monitoring and on-site examination. Off-site monitoring involves analyzing data assembled both from financial statements produced by the bank and from information gathered in the most recent on-site examination. Further, information on local economic conditions, the health of industries to which the bank lends, and market indicators of the bank's own well-being are reviewed. For example, the review may include investigation of local indicators such as bankruptcies, unemployment, housing prices, and vacancy rates (OCC 2001, 6). The review typically involves computer programs that tabulate and produce ratios based on financial data. By providing warning of a bank's declining health prior to a regularly scheduled on-site examination, off-site monitoring can point up the need to quickly conduct an on-site examination. Data from off-site monitoring can also guide examiners to problem areas during an on-site examination.

In principle, the combination of off-site and on-site reviews will reveal a bank's condition. However, in practice, declining bank health is often difficult to assess and involves subjective judgements by examiners. These judgements must be discussed with bank managers and directors, who might be unduly optimistic about the bank's viability. Further, examiner decisions can later be contested in lawsuits. The difficulties of correctly identifying problems were illustrated by the 1999 failure of First National Bank of Keystone, West Virginia. Reportedly, Keystone's management hid the bank's insolvency from banking agency examiners for an extended period directly before examiners closed the bank (Office of Inspector General 2000).

When off-site monitoring or initial indicators from an on-site review raise suspicions that a bank's health may be deteriorating, on-site examiners focus

especially on certain indicators or red flags of declining health. These red flags include especially rapid asset growth, risk management deficiencies, asset quality deterioration, liquidity difficulties, and the bank's unwillingness to cooperate with examiners (OCC 2001, 3–18). During the on-site review of a bank thought to be experiencing trouble, special emphasis will be placed on careful valuation of the bank's assets. For banks that later fail, the valuation often reveals that the bank was unrealistically optimistic when valuing assets in its financial statements.

Following the examination of a seriously troubled bank, examiners will meet with the bank's management and its board of directors, notifying them of negative findings from the examination. Other regulators are also typically notified. For example, state bank examiners are likely to notify the FDIC, or the Federal Reserve if the bank is a Federal Reserve member bank (since the Fed supervises state member banks along with state agencies). If the asset valuation proves that the bank is undercapitalized (capital-to-assets ratio less than 4 percent), and especially if it is found to be critically undercapitalized (ratio below 2 percent), examiners will inform bank management and directors, in writing, of the amount of capital that must be injected to recapitalize the bank. The bank is given the opportunity to produce a plan to gather the necessary capital, and if critically undercapitalized, is warned that it will probably have no more than 90 days in which to gather the needed capital. During this meeting with the bank's board of directors and its management, representatives of the FDIC may be present. They are there to obtain the board's authorization for the FDIC to begin seeking bidders for the bank (OCC 2001, 61–62).

If the bank is successful in raising additional capital, it can continue to operate and avoids closure. On the other hand, if investors are unwilling to provide the needed equity, the bank will be closed and placed into receivership.

6. TREATMENT OF CLOSED BANKS

Since the FDIC not only is receiver, but also deposit insurer, its responsibility surpasses a typical receiver; it must make insured depositors whole. It does so either by repaying depositors out of its insurance reserves or, far more frequently, by transferring the deposits to another bank. The bank that acquires the deposits must be compensated since deposits are simply liabilities representing future payments to be made by the bank acquiring the deposits. Such compensation also is drawn from the FDIC's reserves. The FDIC collects the failed bank's assets by selling them to a healthy bank or other investor, or by holding them and collecting them as they mature. The proceeds of the asset collection are retained by the FDIC, offsetting its insurance losses, and are used to repay other creditors of the failed bank.

The FDIC as Receiver

The FDIC typically employs one of two techniques when acting as the receiver for a failing bank: deposit payoff or purchase and assumption. A deposit payoff involves 1) repaying insured depositors, 2) liquidating assets of the bank, and 3) dividing the proceeds from asset liquidation between itself and uninsured bank creditors. In a purchase and assumption transaction "a healthy institution *purchases* some or all of the assets of a failed bank . . . and *assumes* some or all of the liabilities" (FDIC 1998b, 19). When deciding which of these techniques to employ, the FDIC is guided by the least-cost requirement of the FDICIA. The requirement states that whenever it is named receiver, the FDIC must choose the option that was *least* costly in terms of the FDIC expenditures. The rule was a prominent feature of the FDICIA.[2]

The FDIC's Analysis

The FDIC's role begins when a bank chartering agency determines that the bank is no longer viable and notifies the FDIC. The FDIC then begins a multistep process generally lasting 90 to 100 days, but which can proceed much more quickly. While technically the FDIC is named receiver at the end of the process—on the day the bank is closed—it begins performing many of the activities one expects of a receiver long before this date.[3]

Immediately following the agency's notification of the FDIC, the FDIC performs a careful analysis of the bank's assets and liabilities to estimate the cost the FDIC will incur implementing a deposit payoff. The process involves estimating the market value of the assets. Since there is no secondary market for many loans, the asset valuation is based in part on likely cash flows, discounted to the present, minus costs of holding or selling the loans.

The FDIC also determines the amount of insured deposits, since it is this amount that the FDIC is responsible for repaying. One might imagine then that if the value of assets is smaller than the value of insured deposits, the FDIC simply subtracts the market value of assets from the amount of insured deposits, and the difference is the cost to the FDIC of a deposit payoff. Instead the calculation is somewhat more complicated. The FDIC must divide the proceeds of the asset sale between itself and uninsured depositors. The division is based on the share of total deposits held by insured and uninsured depositors. For example, if insured deposits account for 75 percent of total deposits, then

---

[2] An excellent historical review of the FDIC's work as receiver during the banking and thrift crisis of the 1980s can be found in Volumes 1 and 2 of FDIC (1998a). For a broad historical review of the crisis see Volumes 1 and 2 of FDIC (1997).

[3] See Chapter 2 of FDIC (1998b), which provides a thorough, readable review of the process by which the FDIC closes out the operation of a troubled bank.

the FDIC receives 75 percent of the ultimate proceeds of collecting the assets. Uninsured depositors receive 25 percent (Thomson 1994).

If the value of assets exceeds the amount of total deposits, then the bank's general creditors are next in line to receive payments, followed by subordinated debt holders.[4] Last in line are equity holders. Payments to general creditors, subordinated debt holders, and equity holders only occur if the FDIC and uninsured depositors suffer no losses. Therefore, payments to these creditors and to equity shareholders do not raise the FDIC's cost.[5]

Whether a deposit payoff or purchase and assumption is ultimately chosen, this estimation of the FDIC's cost of deposit payoff gives the FDIC a required baseline as it prepares to sell the bank. No bids can be accepted at a price that yields a closure cost to the FDIC higher than the cost of a deposit payoff because of the least-cost rule.

Bidding

Next the FDIC draws up a detailed description of the assets and liabilities of the troubled bank. Once completed, the FDIC offers the bank for sale, providing the description, known as the *information package*, to interested bidders. Interested bidders are allowed the opportunity to perform a due diligence analysis on the troubled bank, involving on-site review of the bank's books.

Following due diligence, each interested party submits a bid. The bid includes two figures. The first figure is the bidder's estimate of the collectable value of the troubled bank's assets or that portion of the assets that the bidder plans to acquire. When calculating this figure, the bidder deducts his expected collection expenses. The second, the intangible asset portion of the bid, is the bidder's estimate of what is often called the bank's franchise value. Here the bidder is attempting to estimate the future earnings flows that might emanate

[4] There are two scenarios under which a failed bank's assets might exceed deposits. First, the FDICIA authorizes bank regulators to appoint a receiver when a bank's capital falls to 2 percent. In such a case, total assets exceed total liabilities, so assets will certainly exceed deposits which can be no greater than total liabilities. For most banks deposits are significantly smaller than total liabilities (on average deposits account for about 72 percent of bank liabilities). Second, the value of the troubled bank's assets might be smaller than the value of total liabilities, such that the bank has negative capital and must be closed, yet the value of assets exceed the value of all deposits.

[5] The priority of payment is based on a provision of federal law normally referred to as the depositor preference rule, established by Title III of the Omnibus Budget Reconciliation Act of 1993 (Public Law 103–66). Prior to its enactment, general creditors had a claim on the failed bank's assets equivalent to that of the FDIC and uninsured depositors. The law placed the FDIC and uninsured depositors ahead of general creditors. See Thomson 1994 for a review of the changes produced by the depositor preference law. General creditors are all bank creditors except those holding 1) bank deposits or 2) subordinated notes and bonds issued by the bank. Subordinated notes and bonds are bank debts which contractually specify that they are repaid only after other bank creditors are repaid.

from the long-term deposit and loan relationships developed with the failed bank's customers. The bid is based on this estimate (FDIC 1998b, 13–17).

Several factors are important when estimating franchise value. One such factor is the proportion of core deposits in the failed bank's liability portfolio. Core deposits—checking and savings accounts held by individuals and small businesses—tend to pay below-market interest rates. Further, the depositors holding these core deposits are often deemed unlikely to withdraw their funds when a new bank takes over because of the inconvenience and costs of doing so. Acquirers are willing to pay more for a bank with a high proportion of this long-term source of low-cost funding. Another factor likely to be an important determinant of franchise value is the expected economic growth of the bank's market area. The relationships established by a bank with its retail and business customers, whether the bank is healthy or failing, are far more valuable in a vibrant local economy than in one in decline.

While the potential acquirers' bids are made up of two figures, one for the value of assets acquired and one for the franchise value acquired, another figure also plays an important role in the bid amount. Winning bidders typically assume the failing bank's insured deposits, and these deposits often exceed the value of assets purchased. Therefore, the acquirer will not pay the FDIC to buy the bank but instead must, on net, be compensated to take over the failing bank. As a result, one might assume that in total an acquirer's bid would be a negative dollar amount. But by convention, bids are calculated as if the FDIC were to pay the winning bidder dollar-for-dollar to assume insured deposits. For example, consider the case of Comatose National Bank. At the time of its failure the book value of its assets was $200 million, and it had $220 million in insured deposits. The amount bid by the winning bidder, Acquisitive National, was $130 million, its estimate of the collectable value of the assets, and $10 million for franchise value.

Though Acquisitive's bids together summed to $140 million, it did not pay this amount to the FDIC at consummation. Instead, the FDIC paid Acquisitive $80 million ($220 million minus $140 million), since Acquisitive had to be compensated for assuming $220 million in Comatose deposits.

Once the bidding closes, the FDIC accepts the bid that produces the least cost to the FDIC. As long as one of the bids exceeds the FDIC's estimate of the cost of liquidating assets and repaying insured depositors—the deposit payoff method of disposing of the troubled bank—then the FDIC will employ the purchase and assumption method.

Uninsured Depositors Often Lose

Under either the deposit payoff or purchase and assumption method, uninsured depositors and other bank creditors stand to suffer losses. Losses have not always been imposed on such depositors. Prior to the FDICIA the FDIC could choose any resolution with a cost below the cost of a deposit payoff.

As a result, the FDIC had the leeway to make uninsured depositors whole even if doing so was not least cost, allowing it to protect the majority of these depositors from loss. The least-cost rule makes covering uninsured depositors much less likely, so that since the FDICIA, most have suffered losses. Even today, when the least-cost bidder wishes to take the uninsured deposits, uninsured depositors can be protected.

Frequency with Which the Disposition Methods Have Been Used

The FDIC employs three methods to handle failing banks. The first, purchase and assumption transactions, have predominated throughout the FDIC's history, accounting for 66 percent of all transactions since the FDIC was formed (FDIC 2003). During the 1980s, in exchange for taking on certain failing banks, purchasers often benefited from relaxed regulatory treatment. For example, the Garn-St. Germain Act of 1982 authorized the FDIC to sell failing banks to banks located outside the failing bank's home state. At the time, interstate banking was largely prohibited, so the opportunity to move into an attractive state by purchasing a failing bank in that state was especially appealing to acquisitive banks. Acquirers, in turn, were willing to pay more, diminishing the failed bank's drain on the FDIC's reserves, bank insurance premiums, and the chance that taxpayers would be called on to bailout deposit insurance. The second, deposit payoff transactions, have amounted to 20 percent of all FDIC transactions. A third transaction type, open bank assistance, accounts for 6 percent of all FDIC transactions.[6] In an assistance transaction, the FDIC provides cash or loans to an insolvent bank to return it to solvency, thereby preventing its failure.

Congress authorized the FDIC to engage in open bank assistance in 1950, but only if the troubled institution was deemed essential to its community (FDIC 1998b, 48). The FDIC assisted only six banks from 1950 until 1982, when the Garn-St. Germain Depository Institutions Act dropped the essentiality test. Instead, Garn-St. Germain authorized the FDIC to provide open bank assistance as long as doing so was less expensive than a deposit payoff. A payoff is usually quite costly to the FDIC, especially when the failing bank is large, since the FDIC must manage payouts to all insured depositors and affect either the sale or collection of all of the bank's assets, a process that can proceed for years. Between 1982 and 1992 the FDIC provided assistance to 120

[6] A fourth type of transaction, the insured deposit transfer, has also been employed. It is a hybrid between a purchase and assumption and a deposit payoff, making up 8 percent of all FDIC transactions. The FDIC began using the deposit transfer in 1983. Under this method a healthy bank acquires the insured deposits of a failed bank but typically none of its assets. In return for taking on these liabilities, the bank receives a payment of cash from the FDIC. The healthy bank benefits from new deposit customer relationships, and, therefore, the FDIC's payment to the acquiring bank is lower than the FDIC would otherwise pay out to depositors (FDIC 1998b, 45).

banks. But in 1991, the FDICIA restricted the ability of the FDIC to choose to assist banks. The Act required that the FDIC always choose the transaction that was least costly in terms of the FDIC expenditures and, should FDIC reserves be depleted, least costly to taxpayers. Further, in 1993 the Resolution Trust Corporation Completion Act (Public Law 103–204) prohibited the use of any FDIC funds to provide assistance if such assistance would benefit shareholders of the troubled institution. The combination of FDICIA's restriction and the Completion Act's prohibition meant that (except in a special case discussed later) the FDIC could no longer provide open bank assistance. Ultimately, this change meant that Congress reduced the FDIC's discretion to use funds from reserves built by bank-paid premiums or taxpayers' dollars for objectives other than closing banks at minimum cost.

Since 1993, therefore, the FDIC has used only purchase and assumption and depositor payout to dispose of failed bank assets and liabilities. Purchase and assumption dominates since it is typically the least-cost method. It is least costly because a purchaser is normally willing to pay more for assets (or equivalently, require lower compensation to assume liabilities) if they are delivered together with associated liabilities. When a significant portion of a failed bank's assets and liabilities are sold as a bundle, the buyer acquires the benefits of continuing relationships with the failed bank's customers. Developing these relationships could only be accomplished at a significant cost otherwise.

7. WHAT DID THE IMPROVEMENT ACT IMPROVE?

The Federal Deposit Insurance Corporation Improvement Act was intended to ensure that 1) bank closures be accomplished at the minimum cost possible to the deposit insurance fund, and 2) that bank supervisors quickly close banks that are no longer viable. Closure policies whereby uninsured depositors were protected from loss came to be viewed as unacceptably expensive to the insurance fund as well as damping any incentive these depositors might have to monitor their banks. During the 1980s, bank supervisors had been slow to close banks and thrifts that were in serious trouble, betting that the institution would recover if given time. Such delay was viewed by many observers as contributing to considerably higher closure costs and to excessive risk-taking by the troubled banks. Further, acting slowly meant that financial resources were locked away in poorly run operations.

The Act appears to have gone a long way toward accomplishing these two goals. First, as mentioned earlier, the FDICIA has meant that uninsured depositors suffer losses much more frequently. In the years immediately before implementation of the FDICIA, meaning 1986 through 1992, in 78 percent of bank failures uninsured depositors suffered no losses. Since the FDICIA became effective, (i.e., from 1993 through 2002), in 76 percent of bank failures

uninsured depositors were repaid less than 100 percent of the value of their deposits (Benston and Kaufman 1997, Table 3; FDIC *Annual Report*, various years). As a result, these depositors now share some of the bank failure costs previously borne by the FDIC. Uninsured depositors are those with more than the FDIC coverage limit of $100,000 in an account and depositors with funds in foreign offices of U.S. banks.

As an additional advantage, because depositors are no longer protected from loss, those with more than $100,000 in an account or with funds in a bank's foreign office can be expected to exercise some market discipline over banks. In other words, these individuals will make an effort to monitor the health of their banks since they know that depositors in failed banks have experienced losses. They will demand higher interest rates or remove their funds if the bank is perceived as too risky. Bank risk-taking will be reduced, lowering the frequency and cost of bank failures.

The monitoring benefit may be limited, however. The FDICIA appears to have increased the likelihood that large banks' uninsured depositors may suffer losses in a failure, and large banks are the greatest recipients of uninsured deposits. Prior to the FDICIA, these depositors were uniformly protected (Benston and Kaufman 1997, 30). However, in the case of several fairly large banks, depositors have suffered losses since the FDICIA. Yet, FDICIA contains a provision that allows supervisors in some cases to protect all depositors in large banks, a fact which will certainly damp depositors' incentives to monitor large bank health. Further, uninsured deposits make up a fairly small fraction of the total deposits of small banks, so for these banks the significance of depositor monitoring is also limited. As of the end of the first quarter of 2003, on average, uninsured deposits accounted for only 23 percent of small banks' deposits. Here a small bank is defined as one with assets less than $1 billion.

Second, the FDICIA set in place a mechanism to lower the likelihood that supervisors will forbear, i.e., wait to close a bank that is clearly no longer viable. Since 1993, when the prompt corrective action provisions took effect, supervisors generally have had 90 days, and at most 270 days, to close a critically undercapitalized bank. While the requirement is straightforward, measuring a bank's capital—meaning valuing assets and liabilities—often is not. Any failure by supervisors to quickly force a bank to "write down" the value of uncollectible assets will diminish the effectiveness of these prompt corrective action requirements.

Placing a dollar value on assets and liabilities always involves subjective judgements. One of the primary roles of bank supervision is measuring this value. If the bank has overstated capital, examiners from the bank's supervisory agency should require the bank to reduce its reported level of capital, in some cases enough to indicate that the bank is insolvent or nearly so. Ensuring that these examiner judgments are made in an unbiased manner—at times when the management of the troubled bank and interested parties might bring

pressure to be lenient—is critical to the effectiveness of prompt corrective action.

The FDICIA established a mechanism intended to encourage unbiased judgments by bank supervisors. If the supervisor does not lower the value of capital when doing so is appropriate, the troubled bank is likely to increase its losses, ultimately expanding the FDIC's losses when the bank is closed. Whenever the FDIC loses an amount equal to the greater of $25 million or 2 percent of the failed bank's assets, the inspector general of the failed bank's federal supervisor must prepare a report on the failure. The report describes the reasons for the loss and how such losses might be prevented in the future. These reports, scrutinized by the General Accounting Office, are available to Congress and to the public. The threat of public scrutiny and censure provided by an inspector general report is intended to offset any pressures to be inappropriately lenient.

The inspector general reports appear to have had a positive influence on banking agency enforcement of the prompt corrective action portions of the FDICIA. According to Benston and Kaufman (1997, 21, 26) several critical reports were produced soon after the FDICIA requirement became effective. As a result, the agencies modified their procedures and received more favorable reports subsequently.

8. LIMITS ON DISCOUNT WINDOW LENDING

The FDICIA's least-cost requirement successfully changed FDIC behavior so that uninsured depositors are much less frequently protected. Yet, these depositors might still avoid losses, and so have little incentive to monitor. As a bank's financial health deteriorates, depositors may become aware of this deterioration, and those with deposits not protected by the FDIC will seek to withdraw them. The bank might attempt to meet depositors' demands for repayment by borrowing from the Federal Reserve. If the Fed lends, depositors can escape and avoid losses. Consequently, if depositors expect Fed discount window lending to flow, they have a reduced incentive to monitor bank risk-taking.

Further, uninsured depositors' gain is met by an equivalent FDIC loss. Fed lending that allows these depositors to escape, increases FDIC losses because, while uninsured depositors bear losses in a bank failure, the Fed does not. The Fed bears no losses since it lends only if the bank provides, as collateral, assets worth more than the Fed loans. Therefore, funding from depositors who are likely to share some of the failed bank's losses, is replaced by funding from the Fed, which is protected against losses. Losses that might have been borne by depositors are transferred to the FDIC, and the vehicle making this transfer possible is Fed lending.

Legislators designing the FDICIA saw the danger and placed limits on Fed lending. In general, the Act allows lending for only 60 days to banks that are undercapitalized. Even more restrictive is the limitation on lending to critically undercapitalized banks. For such banks, the Fed can only lend for five days. So the FDICIA restricts the potential for Fed lending to allow uninsured depositors to escape.

The Act's limitations may not be perfect, however. For as Broaddus (2000) and Goodfriend and Lacker (1999) note, depositors may escape before the limitations are brought to bear. Depositors may have become aware of financial problems before the bank's reported capital has fallen. In other words, while the bank's true capital position may be quite weak, reported capital levels may not have declined. Examiners may not yet have forced a revaluation of the bank's assets. If so, Fed lending might occur, allowing uninsured depositors to escape from a weak bank.

Yet even if FDICIA's limitations prevent Fed lending from facilitating the escape of depositors, some may still escape. For even without any Fed loans, the bank could meet the demands of depositors and creditors by selling the same assets it would otherwise release to the Fed as collateral. The assets might be sold to other banks or to secondary market participants. The development during the 1990s of secondary markets for a number of bank loans improves the opportunities for such sales. Consequently, regardless of Fed lending, some uninsured depositors can be expected to escape.[7]

9. AID FOR BANKS THAT MIGHT POSE A SYSTEMIC RISK

Section 141 of the Act requires the FDIC to determine and employ the least-costly resolution method. Further, this section of the Act prohibits the FDIC, when acting as receiver for a troubled bank, from protecting uninsured depositors and the bank's other creditors if doing so adds to the expense of resolution. Yet, Section 141 grants an exception to these rules. The Act itself calls this the "systemic risk" exception, and observers typically refer to it as the "too-big-to-fail" exception. The least-cost rule can be bypassed if the FDIC determines that closing the troubled bank without protecting uninsured depositors or creditors would have serious effects on economic conditions or financial stability. In other words, an exception to the rule is allowed if a bank's closure with losses to uninsured depositors might lead to the spread of financial problems widely through the banking system.

[7] A troubled bank might also repay uninsured depositors by raising new insured deposits. Insured depositors will be willing to provide funding to a troubled bank since they are insensitive to the bank's financial condition. Raising new deposits may take some time. The process can be quickened by employing a deposit broker. The broker can offer a bank's insured CDs to a wide audience of potential investors. However, Section 301 of FDICIA restricts financially weak banks' use of brokered deposits.

Only when the FDIC's Board of Directors, the Board of Governors of the Federal Reserve, and the Secretary of the Treasury in consultation with the President agree to the too-big exception is the latter allowed. Any decision to employ the exception must be reviewed by the General Accounting Office. If the exception is granted, the FDIC must recover its losses from a special assessment on insured banks beyond normal deposit insurance premiums.

Because of the possibility that the exception might be invoked and all depositors protected, investors in those banks that might be granted the exception have a reduced incentive to monitor and are likely to charge less than completely risk-adjusted interest rates. In response, such banks are likely to engage in an inefficiently high level of risk-taking, wasting financial resources.

10. CONCLUSION

During the 1980s, regulators were slow to close troubled thrifts, leaving many such institutions open long after they were clearly insolvent. Rather than forcing ineffectual institutions to disgorge their financial assets so that they might be reemployed in more profitable uses, regulators allowed these assets to remain under the institutions' control. In 1991, legislators passed the Federal Deposit Insurance Corporation Improvement Act. The intention was to reduce the likelihood that taxpayers would again be called on to bail out the deposit insurance fund as they had following the thrift failures of the 1980s. But the Act also established a process more likely to produce an efficient allocation of financial assets. Under the regime introduced by FDICIA, a troubled bank's management must either gather new equity funding or be closed. The requirement that it gather new funding forces the bank to face a market test similar to that faced by troubled nonbank firms. If investors can be convinced that the bank is viable, they will advance new equity, and the bank continues to control its assets. If not, the bank is taken over by a receiver, and its assets are sold to others.

The FDICIA cannot guarantee that the mistakes made during the 1980s will not be repeated, but it has implemented a more market-like set of closure procedures. Further, the Act places limits on regulator discretion and establishes disclosure requirements that encourage regulators to move quickly to force the recapitalization or closure of troubled banks. As a result, the events of the 1980s in which numerous insolvent thrifts were allowed to remain open for extended periods are less likely to be repeated. Still, the Act allows exceptions for certain banks, those deemed to pose a systemic risk. In such cases the Act tolerates a process that differs greatly from the one the market imposes on troubled nonbank firms.

Deposit insurance blunts the default mechanism that disciplines prompt closure of nonbank firms. Therefore supervisors must play a critical role ensuring that insolvent banks are closed promptly. Prompt closure is necessary

to minimize distortions to financial markets and to control the cost of deposit insurance to the banking system and ultimately to the public.

REFERENCES

Benston, George J., and George G. Kaufman. 1997. "FDICIA after Five Years: A Review and Evaluation." Working Papers Series. Issues in Financial Regulation. WP-97-1. Federal Reserve Bank of Chicago. Chicago, IL. (July).

Broaddus, J. Alfred, Jr. 2000. "Market Discipline and Fed Lending." Remarks before the Bank Structure Conference Sponsored by the Federal Reserve Bank of Chicago. Chicago, IL. 5 May. Also available online at http://www.rich.frb.org/media/speeches/show.cfm?SpeechID=21 (accessed November 10, 2003).

Federal Deposit Insurance Corporation (FDIC). 1997. *History of the Eighties, Lessons for the Future* 1, 2. Washington, D.C.

_____. 1998a. *Managing the Crisis: The FDIC and RTC Experience* 1, 2. Washington, D.C.

_____. 1998b. "Resolutions Handbook: Methods for Resolving Troubled Financial Institutions in the United States." Washington, D.C. Also available online at http://www.fdic.gov/bank/historical/reshandbook/index.html (accessed November 10, 2003).

_____. 2003. Historical Statistics on Banking Table CB02. "Changes in Number of Institutions, FDIC-Insured Commercial Banks, United States and Other Areas." Washington, D.C. Also available online at www.2.fdic.gov/hsob (accessed November 10, 2003).

_____. Various years. *Annual Report*. Washington, D.C.

Goodfriend, Marvin, and Jeffrey M. Lacker. 1999. "Limited Commitment and Central Bank Lending." Federal Reserve Bank of Richmond *Economic Quarterly* 85 (Fall): 1–27.

Thomson, James B. 1994. "The National Depositor Preference Law." Federal Reserve Bank of Cleveland *Economic Commentary* (15 February).

U.S. Department of the Treasury, Office of the Comptroller of the Currency. (OCC). 2001. "An Examiner's Guide to Problem Bank Identification, Rehabilitation, and Resolution." Washington, D.C. (January). Also available online at www.occ.treas.gov/prbbnkgd.pdf (accessed November 10, 2003).

_____, Office of the Inspector General. 2000. "Material Loss Review of The First National Bank of Keystone." Report OIG-00-067. Washington, D.C. March 10. Also available online at www.treasury.gov/offices/inspector-general/audit-reports/2000/oig00067.pdf (accessed November 10, 2003).