Q3
1999

Economic & Financial Review

Federal Reserve Bank of Dallas

Economic and Financial Review

Federal Reserve Bank of Dallas

Economic and Financial Review, published quarterly by the Federal Reserve Bank of Dallas, presents in-depth information and analysis on monetary, financial, banking, and other economic policy topics. Articles are developed by economists in the Bank's Economic Research and Financial Industry Studies departments. The views expressed are those of the authors and do not necessarily reflect the positions of the Federal Reserve Bank of Dallas or the Federal Reserve System.

Subscriptions are available free of charge. Please direct requests for subscriptions, back issues, and address changes to the Public Affairs Department, Federal Reserve Bank of Dallas, P.O. Box 655906, Dallas, TX 75265-5906; call 214-922-5254; or subscribe via the Internet at www.dallasfed.org. *Economic and Financial Review* and other Bank publications are available on the Bank's web site, www.dallasfed.org.

Contents

What Credit Market Indicators Tell Us

John V. Duca

This article reviews the economic ideas behind certain domestic interest rate spreads and Federal Reserve surveys of bank loan officers.

John V. Duca is a senior economist and assistant vice president in the Research Department at the Federal Reserve Bank of Dallas.

Although the behavior of credit markets has long been recognized as revealing much about the U.S. business cycle, the economic meaning of credit market indicators has changed. In particular, the differences—or spreads—between interest rates on various private- and public-sector debt contain much valuable information, but the economic inferences we can draw from them have not always remained the same. Thus, it is important to interpret different credit market indicators carefully.

Such indicators drew much attention in fall 1998, when financial markets were affected by the global economic crisis and concerns that the United States could face a credit crunch, in which borrowers have trouble obtaining loans or must pay interest rates far above U.S. Treasury rates. At that time, world equity prices plunged, and many U.S. firms found it difficult to issue new credit instruments. And while Treasury rates fell as investors fled to these safe instruments, interest rates on private debt barely declined and in some cases rose. As a result, the spread between interest rates on ten-year Baa-rated corporate and Treasury bonds widened to levels typically seen in recessions (*Figure 1*). In the past, however, movements in this spread have not always been a reliable indicator of business cycle downturns. One reason is that interpreting credit market indicators can be complicated in periods of market turmoil.

This article provides an overview of several credit market indicators, showing that it is important to carefully interpret what they can tell us. The article reviews the economic ideas behind certain domestic interest rate spreads and Federal Reserve surveys of bank loan officers. The historical relationships of these indicators to the U.S. business cycle are briefly assessed and illustrated.[1] This article then interprets what these varied indicators have been telling us about credit market conditions since late summer 1998, when securities markets were very turbulent.

WHAT DO INTEREST RATE SPREADS TELL US?

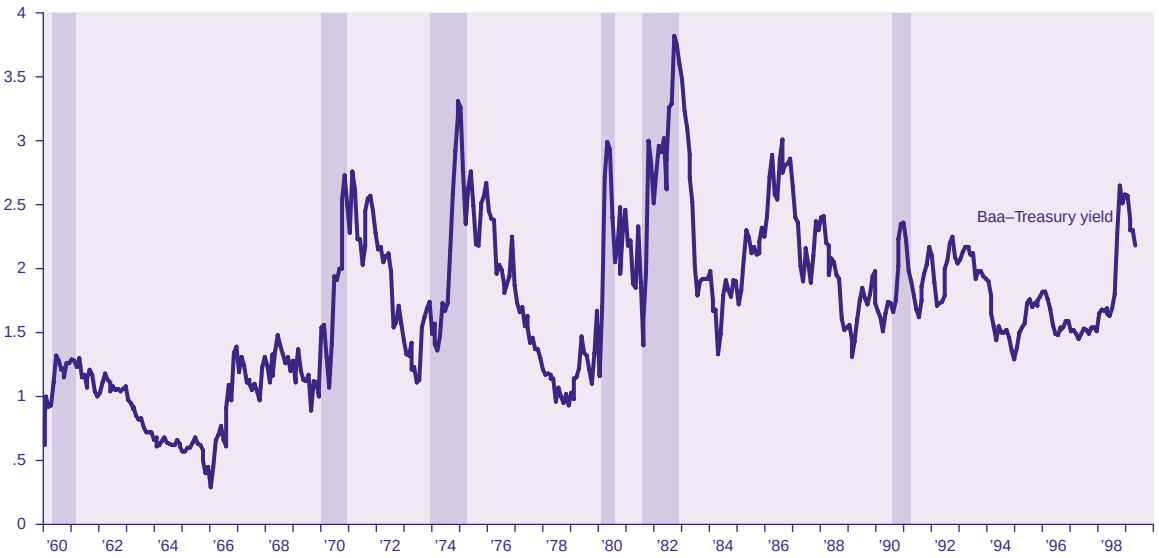
Why It Is Important to Decompose Simple Interest Rate Spreads

The Baa–Treasury spread has risen sharply during or before recessions and even when recessions have not occurred. This mixed record may result from this spread's combination of three different types of risk, for which investors demand compensation and which have had different economic implications.[2]

One component of the Baa–Treasury

Figure 1

Corporate–Treasury Bond Spreads Have Falsely Signaled Recession



NOTE: Bars indicate recessions.

SOURCES: Moody's Investors Service; Federal Reserve Board.

spread is the prepayment premium to investors for the risk that if interest rates fall in the future, borrowers might retire old debt with new debt at lower rates. Because investors can lose in such cases, investors demand an extra return on a bond if the issuer can pay it before maturity. This extra return equals the perceived prepayment risk multiplied by the market price of that risk, both of which can vary.

Another component of the Baa–Treasury spread is a liquidity premium that compensates investors for the fact that private instruments are less desirable to hold relative to U.S. Treasurys when financial markets are turbulent and investors are very risk averse. The Baa–Treasury spread also contains a default risk premium to compensate lenders for the risk that borrowers may not repay, reflecting the amount of default risk posed and the price of risk.

These components of the Baa–Treasury spread have behaved differently and have different implications. For example, at first glance, the widening of the spread between yields on ten-year Baa-rated corporate and Treasury bonds in late 1998 might suggest the risk of an impending recession. However, a less alarming picture emerges from decomposing this spread into the yield spread between Aaa- and Baa-rated bonds, and the yield spread between the highest-grade corporate bond (Aaa) and Treasurys:

(Baa–Treasury) = (Baa–Aaa) + (Aaa–Treasury).

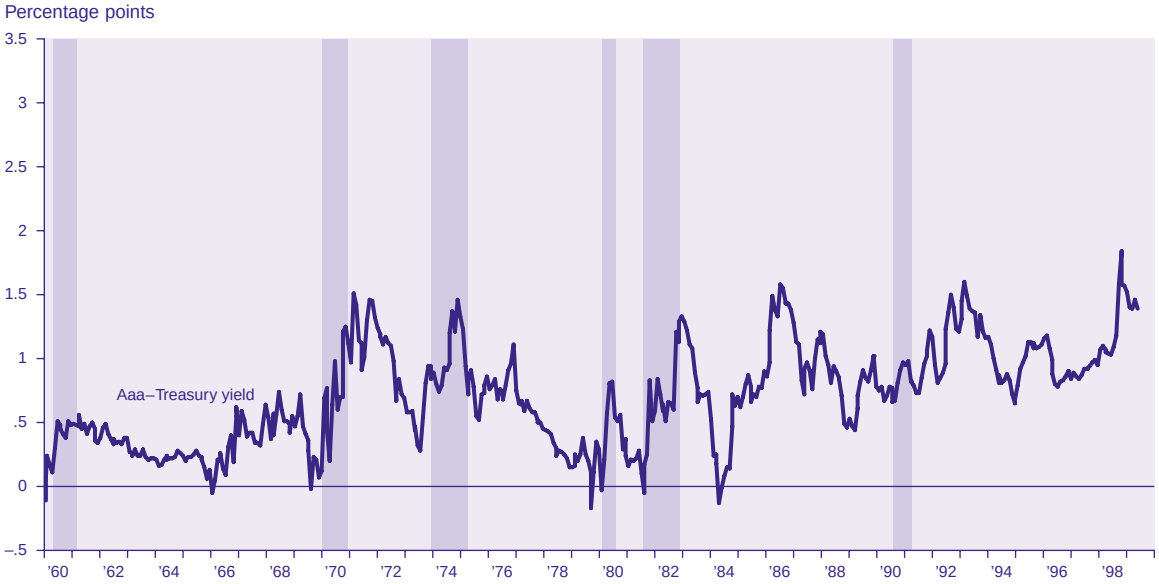
As shown below, the former component (which rose much less) is more reflective of default risk that is correlated with downturns, whereas the latter type of spread (which widened much) is more indicative of prepayment and liquidity risk that is not closely associated with recessions. For these reasons, it is important to interpret different types of interest rate spreads carefully.

Prepayment/Liquidity Premiums in Spreads Between High-Grade Corporate and Treasury Bonds

Investors demand a prepayment risk premium for the possibility that borrowers will refinance their debt if interest rates fall. Under normal conditions, this premium is often tracked by the gap between the average interest rate on callable bonds of the highest grade—Aaa-rated corporate bonds, which pose little default risk—and that on a Treasury bond. With little difference in default risk between such bonds, the primary distinction between Aaa-corporate and Treasury bonds is that when interest rates fall, private-sector bonds often are called and refinanced with new debt, whereas Treasurys are not. For this reason, the Aaa–Treasury spread contains a prepayment risk premium that reflects interest rate risk and the risk of refinancing. However, as noted below, this spread also contains a liquidity premium associated with a more stable demand for Treasury securities.

Figure 2

Higher Prepayment–Liquidity Premiums Often Not Linked to Recessions



NOTE: Bars indicate recessions.

SOURCES: Moody's Investors Service; Federal Reserve Board.

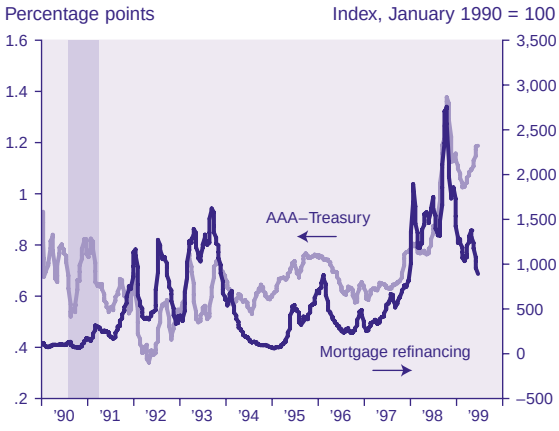
Another complication with yield spreads between corporate and Treasury bonds is that the call provisions on corporate bonds vary within each grade of bond and also change over time. As a result, the prepayment risk premiums in corporate–Treasury yield spreads can vary across time not only due to changes in the risk that market interest rates will vary (henceforth, interest rate risk) but also due to changes in the practices regarding call provisions in bond issues, as emphasized by Duffee's (1998) research. This source of measurement error makes it difficult to separate the time-varying prepayment risk premium in corporate–Treasury yield spreads from other, more economically meaningful components.[3]

One interesting aspect of prepayment spreads is that they are not closely associated with recessions, as shown by the spread between Aaa-rated corporate and Treasury bond yields in Figure 2. This is also true for another measure of prepayment premiums, the interest rate spread between residential mortgage-backed securities and Treasurys. Since mortgage-backed securities are enhanced by collateral and are viewed as having an implicit guarantee from the federal government, these securities are seen as posing little default risk. But, if interest rates fall, many of the securities are retired as the mortgages backing them are refinanced by homeowners. Indeed, the Aaa–Treasury spread and mortgage refinancing activity have swung together since the recession of 1990–91 (*Figure 3*). Nevertheless, it should be noted that changes in the costs and ease of refinancing have affected prepayment premiums over time, as have changes in the liquidity of these securities. This is true of the mortgage-backed securities market, which became deep and well-developed only in the mid- to late 1980s, well after the investment-grade corporate bond market did so.

Figure 3

AAA–Treasury Spread and Mortgage Refinancing Activity Have Swung Together Since the Last Recession

(Four-week moving average)



NOTE: Bar indicates recession.

SOURCES: Moody's Investors Service; Mortgage Bankers Association.

As mentioned above, the Aaa–Treasury spread normally includes not only a prepayment premium but also a liquidity premium that is usually small and less important. However, under unusual circumstances, this liquidity premium can become substantial, compensating investors for the fact that private instruments are less desirable to hold than Treasurys when financial markets are turbulent and investors are extremely risk averse.

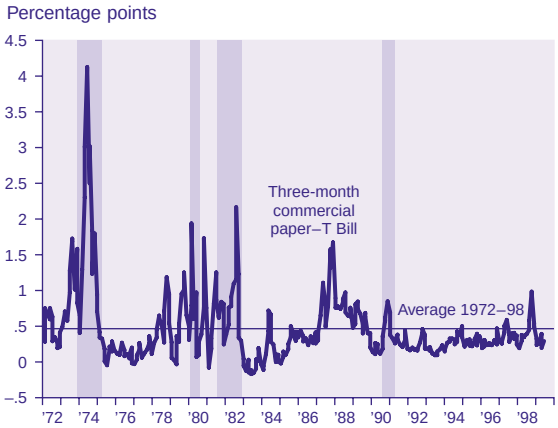
For example, some analysts argued that the big rises in common prepayment premium measures in fall 1998 reflected not so much an increase in prepayment risk as a flight to quality in which investors fled falling stock prices by shifting into the most liquid bond instruments, Treasurys. The flight to quality bid down Treasury yields more than private bond yields, thereby widening the gap between the two. From this point of view, last fall's run-up in the Aaa–Treasury spread is best interpreted as an increase in the so-called liquidity premium associated with a rise in nervousness among investors.

Commercial Paper–Treasury Bill Spreads as Indicators of Liquidity Risk

A clearer gauge of the liquidity premium is the paper–bill spread, the gap between interest rates on top-grade commercial paper and Treasury bills. Since commercial paper is short-lived, it poses virtually no prepayment risk. In addition, because only the most creditworthy companies have enough market credibility to issue these short-term instruments, top-grade commercial paper normally poses little default risk.

At one time, the paper–bill spread was correlated with recessions, as emphasized by Friedman and Kuttner (1992) and Bernanke (1990). However, since the mid-1980s, this spread, like the prepayment spread, has not been closely related to recessions (*Figure 4*). The reasons for the earlier correlation are not completely clear, despite some attempts to explain them (most notably, Bernanke and Blinder 1992). One possibility is that the paper–bill spread spikes during periods of uncertainty, when even the strongest companies posed some default risk. Indeed, Hafer and Kutan (1992) and Emery (1996) found that most of the statistical significance of the paper–bill spread in samples from the 1960s–1980s was the result of an unusual spike in late 1973. This event coincided with the first OPEC crisis of 1973, when an oil embargo hurt the U.S. economy. The OPEC crisis of 1973–74 was the first major supply shock or stagflationary period in decades.

Figure 4
Paper–Bill Spread a Less Reliable Indicator After the Mid-1980s



NOTE: Bars indicate recessions.

SOURCE: Federal Reserve Board.

Because of its unusual character, the jump in both inflation and unemployment confounded many analysts and created uncertainty that led investors to demand large risk premiums. On these grounds, some critics of the paper–bill spread believe that the spike of 1973 reflected the impact of a big supply shock and this coincidence makes the paper–bill spread appear to be a better leading indicator than it really is.

Explanations for the more recent decline in the information content of the paper–bill spread relate to asset substitutability, as stressed in the recent work of Friedman and Kuttner (1998). In practice, commercial paper is highly substitutable for uninsured large time deposits (often called certificates of deposit, or CDs) issued by banks or thrifts. During the thrift crisis of the late 1980s, investors demanded higher yields on many CDs, which because of market practices also drove up commercial paper rates. As a result, the paper–bill spread rose to high levels in 1987 and falsely signaled an impending recession in 1988–89.

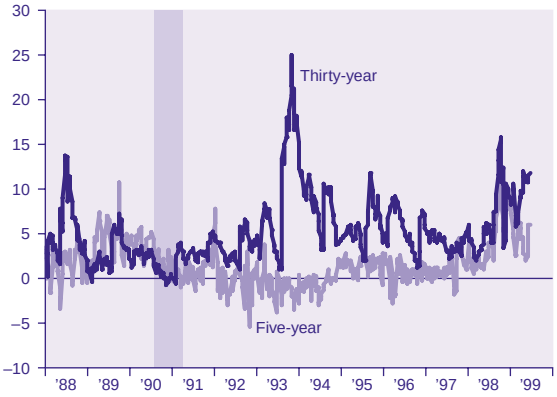
Liquidity Premiums and On-the-Run/Off-the-Run Treasury Spreads

Other indicators of liquidity premiums are on-the-run/off-the-run Treasury yield spreads. These spreads are based on the implied holding-period yields of Treasury securities whose remaining maturities do not precisely match up with those on more recently issued Treasurys. For example, an on-the-run Treasury at a three-year maturity could be the most recently issued three-year Treasury note, while a comparable off-the-run security could be a three-year Treasury

Figure 5

Treasury Liquidity Spreads Are Noisy

Basis points,
on-the-run/off-the-run premiums



NOTE: Bar indicates recession.

SOURCE: Federal Reserve Board.

note issued just a few months earlier. Normally the implied yields on both securities over the next two and three-quarters years would be within a few basis points, with the on-the-run issue having a lower yield. The most recent issue is more liquid, partly because its maturity more closely tracks time in rounded units. In addition, new issues are more liquid with better known trading prices because they have recently been purchased by investors from primary dealers who bid on the bulk of government debt at Treasury auctions.

In turbulent markets, investors could prefer the more liquid on-the-run issue, causing a widening of the on-the-run/off-the-run spread. As shown in Figure 5, such spreads have sometimes surged in times of market uncertainty, such as in late 1989 and late 1998, when stock prices fell. These spreads are indicative of the liquidity of securities markets.

However, it is unclear what relationship these spreads have to the overall U.S. economy. One reason is the data needed to measure such spreads have been consistently saved only since 1987. Because the data span only one business cycle, there is not enough time series evidence to confidently estimate the economic significance of movements in this spread. Another drawback is that on-the-run/off-the-run spreads have been very noisy, sometimes widening during periods of strong GDP growth. Finally, the development of computer-driven trading may have altered the behavior of these spreads and their economic implications over time. For example, some risky investors, including some hedge funds, would bet these spreads would return to normal after widening. Under normal conditions, such strategies would help stabilize these spreads. However, if investors become averse to liquidity risk and wide spreads persist, these strategies can lead to big losses, as happened to a prominent hedge fund in fall 1998.

Default Risk Premiums in Yield Spreads Across Corporate Bond Categories

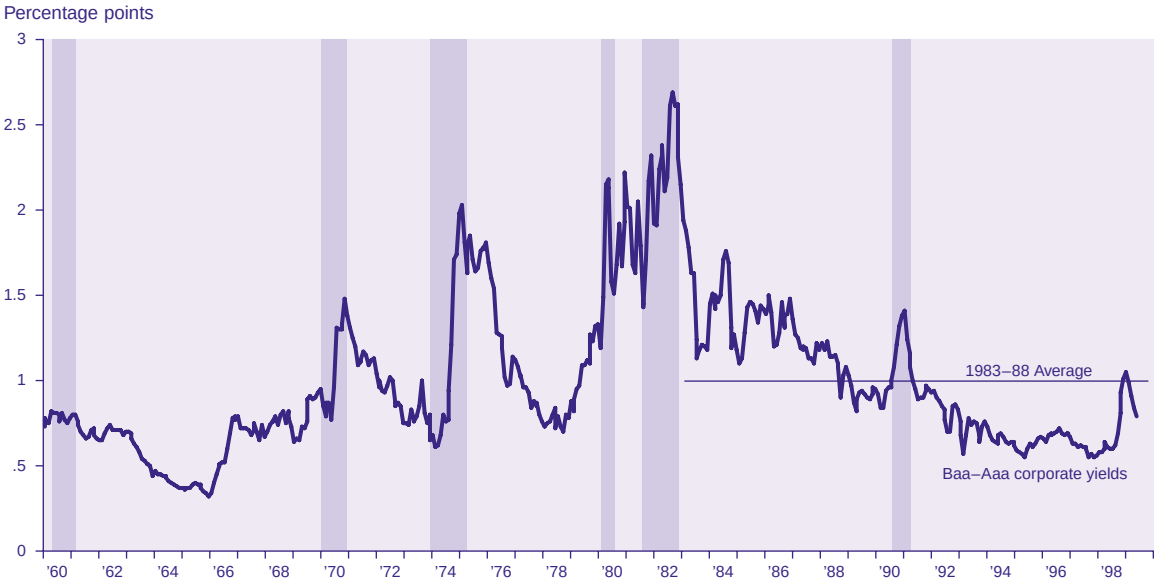
Looking at spreads across different corporate bond categories is advantageous. These corporate spreads are subject to fewer complications posed by prepayment risk than are corporate–Treasury spreads because corporate bonds have similar callability provisions.[4] This implies that such spreads largely reflect default risk premiums. Such premiums, which compensate lenders for the risk that borrowers may not pay back their debt, reflect the market's assessment of the magnitude of default risk posed and the market price of a given amount of risk. The latter depends on the supply of funds to that sector, which in turn depends on how risk-averse and liquid investors are.

Since the amount of default risk and its price reflect cyclical conditions, the spread between Baa and Aaa corporate yields has risen during recessions (*Figure 6*) and—relative to corporate–Treasury spreads—has a much lower tendency to falsely signal recessions. As Jaffee (1975) notes, corporate spreads are significantly related to macroeconomic conditions, both in a statistical and economically meaningful sense. In practice, increases in such premiums also have been associated with a tightening of credit standards, which makes it more likely that credit applicants get turned down by banks or get shut out of the bond market.

There are, however, two drawbacks to using these spreads. First, rather than giving advance warning of recessions, they tend to rise during recessions. This suggests they are better coincident indicators than leading indicators of economic activity. Second, the spreads have tended to decline since 1983, making it difficult to detect recession risk from the level of this spread. For example, the Baa–Aaa spread rose in the 1990–91 recession to a level near the average for the nonrecession months of the 1970s and 1980s. Another recent example is the run-up in this spread during the fourth quarter of 1998, when securities markets were turbulent. By itself, the increase in the spread suggests a rise in the default risk premium. However, because the level rose to the average of the post-1982 period, it is difficult to tell whether the recent run-up reflects a serious risk of recession or a return to more normal risk-

Figure 6

Default Premiums Tend to Rise During Recessions but Their Levels May Have Different Implications Across Time



NOTE: Bars indicate recessions.

SOURCE: Moody's Investors Service.

taking by investors after unusually low spreads during the mid-1990s.

The downtrend of investment-grade spreads since the early 1980s can be largely attributed to a more stable environment stemming from a shift to low inflation and a perception that the U.S. economy is less susceptible to large downturns (for example, see Dudley and McKelvey 1998). Also contributing to the downtrend in these spreads are several factors that deepened the corporate bond market, making it less subject to price volatility associated with thin trading or periods of rumor-driven trading. One factor is the improvement in information technology that has made it easier and cheaper for investors to monitor firms, thereby making investments less uncertain. Other factors have boosted the retirement demand for corporate bond investments, including the aging of the baby boom generation,[5] the post–World War II rise in the overall share of workers having some form of pension benefits, and legal changes fostering the growth of IRAs and 401K defined-contribution pension accounts.[6] Together these factors have made investors more willing to purchase lower rated investment-grade bonds, thereby pushing down spreads such as the Baa–Aaa yield spread.

It is important to note that the Baa–Aaa spread reflects credit market conditions for well-established, highly rated firms, whereas spreads between investment-grade and below-investment-grade bonds (so-called junk spreads) are indicative of credit market conditions for mid-sized, less well established firms. The advantage to using investment-grade spreads is that their data extend far back in time, giving us a record spanning several business cycles. By contrast, junk bond indexes only extend to the mid-1980s, when the junk bond market developed. In addition, the greater liquidity of the investment-grade market implies that these spreads are more indicative of fundamental factors affecting default risk premiums and less indicative of temporary fluctuations due to market turbulence.[7]

Indeed, as shown in Figure 7, junk bond spreads jumped much more during the 1990–91 recession than did the investment-grade spreads shown in Figure 6. The rise in junk spreads during the most recent recession strongly suggests that default risk affects junk spreads. However, variation in liquidity risk plausibly affects junk spreads more than it affects investment-grade spreads. For example, during the early 1990s, junk spreads were boosted by new regulations and the resolution of the thrift crisis, which forced the thrift institutions to sell their junk bond holdings.[8] Given that many important institutional investors could not readily purchase these securities, these sales greatly depressed the prices of junk bonds, thereby pushing up junk bond yields and spreads.

What Are Brady Bond Spreads and What Do They Tell Us?

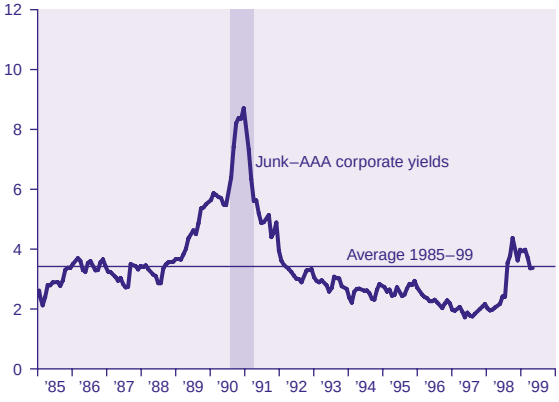
Brady bond interest rate spreads are helpful gauges of credit market conditions in emerging market economies. For reasons specified below, Brady bond spreads largely reflect default risk and are associated with the availability of international funds to emerging economies. What are Brady bonds? In exchange for forgiving many nonperforming loans in the 1980s, lenders were repaid by some emerging market countries with Brady bonds that the lenders could hold in portfolio or sell in credit markets. There are many types of Brady bonds, but all offer some guarantee on the interest payments or principal that removes much, but not all, of their risk. Many guarantees use Treasury bonds as collateral that investors can claim to cover missed interest or principal payments.

To allow better comparisons of Brady bonds with the bonds issued by other governments, the investment industry has created claims on these bonds that take into account these various kinds of partial guarantees. The spreads between the yields on these "stripped Brady bonds" and the yields on Treasury bonds reflect the extra default and liquidity risk that stripped Brady bonds pose relative to the debt of very creditworthy nations such as the United States. While in principle Latin American issuers of Brady bonds can call their debt, in contrast to the typical practice of the U.S. government, Brady bonds are viewed as posing little, if any, prepayment risk.[9]

Figure 7

Junk Bond Spreads Jump to Slightly Above Normal in 1998

Percentage points



NOTE: Bar indicates recession.

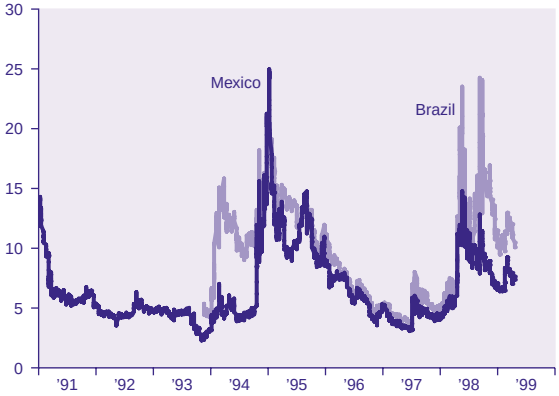
SOURCES: Moody's Investors Service; Merrill Lynch.

Figure 8

Latin Brady Bond Default Risk Erupts Again

(Stripped Brady Bond spreads relative to Treasury yields)

Percentage points



SOURCE: Federal Reserve Board.

The first example of how stripped Brady bond yield spreads behaved in a debt crisis occurred during the 1995 peso crisis, when Mexico devalued the peso after it could no longer defend its fixed exchange rate. The peso's fall made it more uncertain whether Mexican firms and the Mexican government could repay debt for two reasons. First, it implied that Mexican debtors would have to pay more pesos to repay their foreign-denominated debt. Second, the associated decline in the Mexican economy decreased the likelihood that Mexican debtors would have the revenue to repay debt.

As a result, the default risk on Mexican Brady bonds rose, and because the market for them was thin, their liquidity risk premiums also jumped. Compounding these problems was the maturing of short-term debt issued before the crisis. Investors either refused to buy any new debt issued by Mexican firms and the Mexican government to refinance the maturing debt or demanded high interest rates to do so.

Several other Brady bond issuers had followed policies similar to Mexico's, such as fixing exchange rates and borrowing much short-term debt denominated in foreign currencies. Given these similar risks, investors demanded higher yields on Brady bonds or any debt issued in such nations. As a result, stripped Brady bond yields surged as Latin America experienced an international credit crunch—credit inflows that had funded economic growth suddenly dried up while principal payments on old debt flowed out to foreign investors (*Figure 8*). This credit crunch created an economic slowdown in these countries following the peso crisis.

WHAT DO FEDERAL RESERVE SURVEYS OF BANK LOAN OFFICERS TELL US?

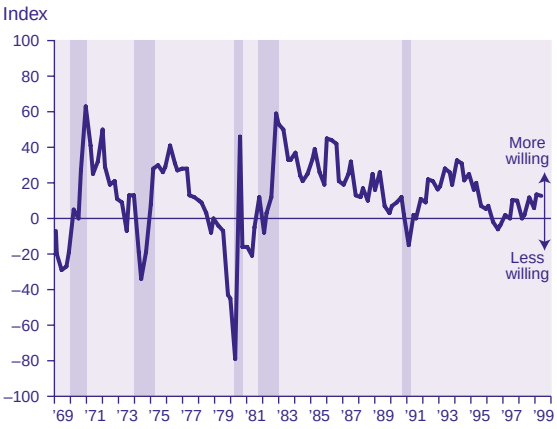
Drawing economic inferences from interest rate spreads is complicated by noise in interest rates and the fact that many such spreads contain different risk premiums that have different economic implications. Therefore, it can be helpful to consider information from surveys of lenders to corroborate evidence on the availability of credit from interest rate spreads. Fortunately, the Federal Reserve has collected such information for three decades.

Specifically, the Federal Reserve has surveyed large U.S. banks quarterly since the late 1960s about their lending practices, conducting up to two extra surveys a year if conditions warrant. The questions have varied over the years, but two types of questions have focused on the degree to which loan applicants have been denied credit. Up until the early 1980s and since the early 1990s, banks have been asked if they have tightened credit standards on business loans and, since the early 1990s, on commercial real estate loans. The quarterly surveys have always asked banks whether they were more or less willing to make consumer installment loans than they were three months earlier. Responses to these two kinds of credit-rationing questions have been particularly informative during credit crunches.

From both types of questions, analysts have created diffusion indexes of the percentage of respondents tightening minus those easing credit standards to see whether large banks had, on net, tightened or eased credit standards. Intuitively, if firms and households are more likely to be denied credit, spending on demand for goods financed with credit would be restrained.

Figure 9

Banks Still Slightly More Willing to Lend to Consumers in 1998–99



NOTE: Bars indicate recessions.

SOURCE: Federal Reserve Board.

Schreft and Owens (1991) show that banks tightened their credit standards on business loans shortly before the recessions of 1970–71 and 1980 but tightened their standards during the recessions of 1974–75, 1981–1982, and 1990–91. These simple patterns suggest that tight credit conditions or credit crunches for businesses may have induced or propagated recessions, respectively. Unfortunately, changes in survey questions in the late 1970s and the absence of such credit-rationing questions during much of the 1980s make it nearly impossible to use this survey evidence to consistently estimate the economic effects of business credit availability over the last few decades.

However, since the late 1960s banks have been asked how their willingness to make consumer installment loans has changed from three months earlier. Using a diffusion index based on this question in econometric models, Duca (1987) and Duca and Garrett (1995) have found that banks' decreased willingness to lend to consumers has a statistically significant negative effect on consumer durable purchases. In addition, this index turned down before most recessions in the United States since the late 1960s, with the notable exception of the last recession, which was arguably prompted by an unexpected disruption of oil markets from the Iraqi occupation of Kuwait (*Figure 9*). The studies mentioned above find that bank willingness to make consumer loans falls as inflation-adjusted interest rates rise or as the economic outlook weakens. Both results support theoretical explanations for the nonprice rationing of credit (see the box titled "Why Loans Are Rationed With Price and Nonprice Terms of Credit"). A decreased willingness to lend to consumers is likely correlated with a tightening of bank credit standards on consumer loans, as implied by Schreft and Owens (1991), who find that movements in diffusion indexes of bank unwillingness to make business loans and tighter credit standards on business loans had a very high correlation (0.80). For these reasons, the index of bank willingness to make consumer installment loans provides an historically long and useful gauge of consumer credit markets.

WHAT HAVE CREDIT INDICATORS TOLD US ABOUT CREDIT CONDITIONS SINCE FALL 1998?

Domestic Interest Rate Spreads

In early fall 1998, financial markets were wracked by turmoil as investors feared that an

Why Loans Are Rationed with Price and Nonprice Terms of Credit

Loans are made using more than just the price of credit (the interest rate) because borrowers may not repay. Typically, lenders offer credit at different interest rates to borrowers posing different levels of default risk, with some applicants denied credit altogether. Assessments are often based on the borrower's credit history, wealth, income, proposed debt payments-to-income ratios, and, for mortgages, down payment ratios (see Duca and Rosenthal 1991; Rosenthal, Duca, and Gabriel 1991). Thus, credit is allocated or rationed using price (loan interest rates and fees) and nonprice terms of credit, both of which can vary.

What could cause such a tightening of nonprice terms of credit? One theoretical approach, typified by Jaffee and Russell (1976) and Stiglitz and Weiss (1981), stresses that lenders bear the downside risks of a loan and face asymmetric information because potential borrowers know more about whether they will repay a loan than do lenders. As interest rates rise, less risky and lower return projects drop out of the pool of loan applications, while riskier ones remain. In addition to this adverse selection effect, there is a moral-hazard problem in that borrowers have more incentive to take bigger risks once they have a loan if they believe they cannot otherwise repay. For some observably risky loan applicants, charging higher loan rates actually worsens loan quality so much that it is not profitable to lend to them. Thus, higher market interest rates or a deteriorating economic outlook makes it unprofitable to lend to what had been marginally creditworthy loan applicants and induces lenders to tighten credit standards used to approve loan applications.

Another approach to explaining the nonprice rationing of credit, typified by Williamson (1986), stresses that lenders bear deadweight costs of default that borrowers do not. These so-called agency costs of default include factors such as legal actions and the interest costs of delays in collection, as well as the time and expense incurred by lending staffs in monitoring delinquent loans and verifying defaults. As stressed by the theoretical work of Townsend (1979), Lacker and Weinberg (1989), and Lacker (1991), debt contracts may be superior to equity contracts for many types of financing. The intuition is that if good economic conditions prevail, borrowers usually meet preset debt payments, and lenders avoid agency costs of verifying how well a firm or household is doing. But, as argued by Bernanke and Gertler (1989), although collateral can reduce this type of agency-cost-induced credit-rationing, declines in asset values brought about by a deteriorating economic outlook or higher interest rates can destroy collateral and cause a tightening of credit standards. Higher interest rates also make it more likely that borrowers will not repay, boosting expected agency costs and prompting tighter credit standards.

economic slowdown would spread from some emerging market economies to the rest of the world. U.S. financial markets appeared to seize up and stop normal functioning. Stock prices were falling sharply and many firms could not issue bonds, commercial paper, or stock. This financial market distress was evident in spreads between corporate and U.S. Treasury bond yields, which jumped sharply. Close examination of the components of such spreads suggests the rises primarily reflected jumps in liquidity and prepayment premiums, as indicated by bigger increases in the Aaa–Treasury spread (*Figure 2*) than in the Aaa–Baa spread (*Figure 6*). Junk spreads widened much more than the Aaa–Baa spread did, but this may have reflected more liquidity than default risk, given the thinness of the junk bond market. Increased prepayment risk was manifested in record levels of mortgage refinancing and a fall in Treasury interest rates. But a flight to quality may have played a bigger role, as evidenced by investors' flight from equities and by a rise in the paper–bill spread (*Figure 5*) and in on-the-run/off-the-run Treasury spreads.

When carefully interpreted, these spreads did not collectively point to recession but, rather, to a scenario of slow growth. Earlier in 1998, fears of slowing export growth from weakening foreign economies led to a decline in both Treasury and private bond rates. This, in turn, stimulated U.S. domestic demand and cushioned U.S. economic growth from a fall in exports (see Duca, Gould, and Taylor 1998). While fears of further global slowing in fall 1998 also sparked declines in U.S. Treasury rates, many private bond rates barely budged. In this sense, the widening of prepayment/liquidity premiums suggested that falling bond yields would not stimulate domestic demand enough to prevent falling net exports from slowing the U.S. economy too much. Against this backdrop and to help stabilize shaky international financial markets, the Federal Reserve cut the federal funds rate three times.

These actions helped restore financial market confidence, as did a natural bounceback in spending that followed a pause in consumption associated with the stock market correction in fall 1998. Since then, the run-up in the paper–bill spread has unwound, while most of the jumps in prepayment/liquidity risk measures have reversed. One interesting development was a further increase and then flattening of the Aaa–Baa spread. Together with other spreads, the rise in this default risk premium to its average level over 1983–98 suggests that while markets are more composed now than last fall, investors are returning to more normal levels of risk-taking in the bond market following the exceptionally easy period of 1996–97.

Foreign Rate Spreads

Brady Bond spreads jumped in fall 1998 to levels not seen since the 1995 peso crisis (*Figure 8*), illustrating investor concerns that emerging market nations would have greater difficulty paying their debts because their economies would slow and currency declines would make it harder for them to pay back in dollars. Since then, spreads have subsided and by May 1999 had indicated that the severe credit crunch gripping Latin America may be lifting. Similar spreads between Asian issues of dollar-denominated bonds and U.S. Treasurys suggest the Asian credit crunch is subsiding.

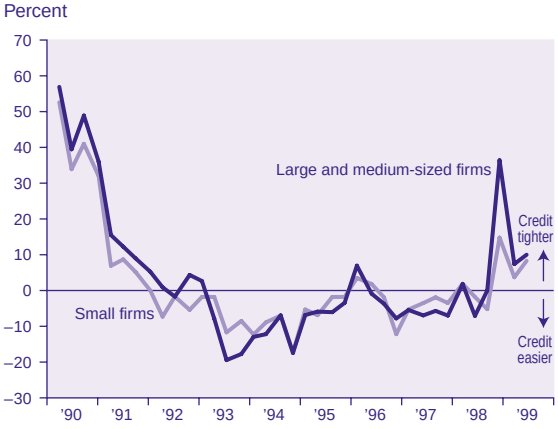
Loan Surveys

The Federal Reserve conducted an extra loan survey in September 1998 that focused on

credit standards. One key finding was that after years of easing credit standards, banks slightly tightened them for loans to large and medium-sized firms, as shown in Figure 10. By contrast, standards were little changed for small firms. Banks that tightened standards did so mainly based on a changing economic outlook. In addition, the larger banks in this sample tended to tighten more than the smaller ones. However, banks reported that they continued, on net, to be more willing to make consumer loans (*Figure 9*). Although the index was less positive, it remained above the negative levels of previous recessions and credit crunches. Other questions revealed slower loan demand by firms and households in September 1998.

The loan surveys have several implications. First, they suggest that credit standards had initially been tightened more for firms with higher global exposure in fall 1998, as such firms usually are bigger and also borrow from larger and more internationally oriented banks. Subsequently, credit standards for large and medium-sized firms have been tightened somewhat further. Second, the surveys imply that small firms experienced a mild tightening of credit standards, but by no means a credit crunch, as confirmed by survey evidence from the National Federation of Small Businesses showing that credit was widely available to the small firms surveyed. Third, while bank willingness to make consumer installment loans has not been increasing as rapidly, households have not been experiencing a credit crunch (*Figure 9*), consistent with strong growth in consumption and consumer credit in late 1998 and early 1999. Together, these three findings suggest that after years of easing standards, lending practices may be returning toward more normal levels of risk-taking. Finally, businesses and household borrowers initially became more cautious, consistent with evidence of little hiring and little firing in early fall 1998. Since then, lending, confidence, spending, and hiring have rebounded, as the caution associated with financial market turmoil has subsided.

Figure 10
Net Percent of Banks Tightening Standards for Business Loans



SOURCE: Federal Reserve Board.

CONCLUSION

This article shows that interest rate spreads and loan surveys should be interpreted carefully when assessing the availability of credit and its impact on the economy. This is especially true of interest rate spread indicators, some of which reflect prepayment, liquidity, or default risk premiums that have different relationships with economic activity. It can be helpful to decompose spreads before drawing economic inferences from the structure of interest rates. Spreads between yields on non-top-grade private-sector bonds and Treasury bonds, in particular, contain a large prepayment premium in addition to a time-varying default risk premium. With respect to recent developments, this distinction especially applies to the Baa–Treasury spread, which could be easily misread as pointing to a severe credit crunch in late 1998. In fact, a simple decomposition of this spread, consistent with other indicators, suggests the United States experienced more of a credit pinch than a credit crunch in late 1998 and early 1999.

It is also important to recognize that even some decomposed spreads contain more than one type of risk premium. In this regard, a widening of some yield spreads that contain a small default risk component, such as the Aaa–Treasury spread, could arise from an increase in prepayment and/or liquidity risk premiums, whose magnitudes may be hard to identify separately. Such was the case in late 1998, when mortgage prepayment activity set records and the commercial paper–Treasury bill rate spread pointed to a jump in liquidity risk premiums.

NOTES

I thank William Gruben and Robert Moore for helpful suggestions and Ricardo Llaudes for excellent research assistance. Any errors are my own.

[1] This article does not review yield curve interest rate spreads, a subject that requires too much space and that has been covered in a host of articles, such as Bernanke (1990). In recent years, the yield curve has

severely underestimated economic growth and has given false signals.

[2] Another component is tax treatment. In contrast to corporate bond interest payments, interest earned on U.S. Treasurys is not subject to state and local taxes. This effect is not likely to shift the spreads significantly because state income tax rates have not varied much over time and because residents in states with high state income taxes tend to buy municipal securities, whereas Treasurys are more likely to be owned by residents in low-state-income-tax states, retirees in low brackets, and institutional investors, who avoid most income taxation. An additional complication is that even if state income tax rates did not vary, the tax-induced spread between corporates and Treasurys could vary because the value of the tax exemption is proportional to the level of interest rates.

[3] There are other complications as well. For example, the Aaa–Treasury spread has even turned negative on rare occasions when the yield curve was inverted after steep rises in interest rates. During these episodes, markets may have anticipated such little prepayment risk that the expected lifetime of corporate bonds exceeded that of the noncallable ten-year Treasury note. (Many corporate bonds have stated maturities greater than ten years.) The implied negative term premium apparently outweighed the liquidity and default risk advantages of the ten-year Treasury note. This was so much the case that the ten-year Treasury yield exceeded the average yield on Aaa-rated corporate bonds, but the latter still exceeded the thirty-year Treasury bond yield in such yield-curve twists when the thirty-year Treasury bond was available.

[4] Nevertheless, yield spreads across corporate bond categories may reflect some differences in callability provisions across corporate bonds, as suggested by Duffee's (1998) study.

[5] Recent cross-section data on households indicate a general shift in household portfolios toward bond and equity mutual funds, for both IRA/401K assets and non-IRA/401K assets (see Kennickell and Starr-McCluer 1994).

[6] In general, pensions include traditional defined-benefit plans and IRA and 401K plans. Since the 1970s there has been a shift away from defined-benefit pension plans and toward defined-contribution pension plans. One advantage of the latter is that a greater share of the expected benefits is portable if employment at a particular firm ends.

Gustman and Steinmeier (1992) and Ippolito (1995) estimate that half the rise in the share of defined-contribution plans (401K and traditional defined-contribution plans as a share of primary pension plans) results from employment shifts away from firms that historically have favored defined-benefit plans—particularly unionized and larger firms. Ippolito (1995) concludes that the other half of this rise stems from tax law changes that made 401K plans more attractive than pre-1980 defined-contribution plans.

[7] The investment-grade market is more liquid because many institutional investors (such as pension funds and life insurance companies) are explicitly or implicitly prohibited from investing in below-investment-grade bonds.

[8] I am indebted to Harvey Rosenblum for pointing this out to me.

[9] There are two primary reasons. First, most Brady bond issuers are unlikely to be in a position to run budget surpluses to pay down debt earlier than scheduled. Second, these nations are unlikely to refinance Brady bonds with new debt having lower interest rates because the original Brady bonds were issued at low rates with collateral backing from major industrialized nations.

REFERENCES

Bernanke, Ben S. (1990), "On the Predictive Power of Interest Rates and Interest Rate Spreads," *New England Economic Review* (November/December): 51–68.

Bernanke, Ben S., and Alan S. Blinder (1992), "The Federal Funds Rate and the Channels of Monetary Transmission," *American Economic Review* 82 (September): 901–21.

Bernanke, Ben S., and Mark Gertler (1989), "Agency Costs, Net Worth, and Business Fluctuations," *American Economic Review* 79 (March): 14–31.

Duca, John V. (1987), "The Effects of Credit Availability on Consumer Durable Expenditures," Federal Reserve Board Economic Activity Section Working Paper no. 80, (Washington, D.C., November).

Duca, John V., and Bonnie Garrett (1995), "Credit Availability, Bank Consumer Lending, and Consumer Durables," Federal Reserve Bank of Dallas Research Paper no. 9514 (Dallas, October).

Duca, John V., David M. Gould, and Lori L. Taylor (1998), "What Does the Asian Crisis Mean for the U.S. Economy?" Federal Reserve Bank of Dallas *Southwest Economy*, Issue 2, March/April, 1–6.

Duca, John V., and Stuart S. Rosenthal (1991), "An Empirical Test of Credit Rationing in the Mortgage Market," *Journal of Urban Economics* 29 (March): 218–34.

Dudley, William C., and Edward F. McKelvey (1998), "The Brave New Business Cycle: Seven Years Old and Still Counting," *U.S. Economic Research*, (March 17), Goldman Sachs Global Economic Research.

Duffee, Gregory R. (1998), "The Relation Between Treasury Yields and Corporate Bond Yield Spreads," *Journal of Finance* 53 (December): 2,225–41.

Emery, Kenneth M. (1996), "The Information Content of the Paper–Bill Spread," *Journal of Economics and Business* 48 (February): 1–10.

Friedman, Benjamin, and Kenneth Kuttner (1992), "Money, Income, Prices, and Interest Rates," *American Economic Review* 82 (June): 472–92.

——— (1998), "Indicator Properties of the Paper–Bill Spread: Lessons from Recent Experience," *Review of Economics and Statistics* 80 (February): 34–44.

Gustman, Alan L., and Thomas L. Steinmeier (1992), "The Stampede Toward Defined Contribution Pension Plans: Fact or Fiction?" *Industrial Relations* 31 (Spring): 361–69.

Hafer, R. W., and Kutan, A. M. (1992), "On the Money-Income Results of Friedman and Kuttner," Southern Illinois University at Edwardsville Working Paper no. 92–03 (Edwardsville, Ill.).

Ippolito, Richard A. (1995), "Toward Explaining the Growth of Defined Contribution Plans," *Industrial Relations* 34 (January): 1–20.

Jaffee, Dwight M. (1975), "Cyclical Variations in the Risk Structure of Interest Rates," *Journal of Monetary Economics* 1 (July): 309–25.

Jaffee, Dwight M., and Thomas Russell (1976), "Imperfect Information, Uncertainty, and Credit Rationing," *Quarterly Journal of Economics* 90 (November): 651–66.

Kennickell, Arthur B., and Martha Starr-McCluer (1994), "Changes in Family Finances from 1989 to 1992: Evidence from the Survey of Consumer Finances," *Federal Reserve Bulletin* 80 (October): 861–82.

Lacker, Jeffrey M. (1991), "Why Is There Debt?" Federal Reserve Bank of Richmond *Economic Review,* July/August, 3–19.

Lacker, Jeffrey M., and John A. Weinberg (1989), "Optimal Contracts Under Costly State Verification," *Journal of Political Economy* 97 (December): 1,345–63.

Rosenthal, Stuart S., John V. Duca, and Stuart A. Gabriel (1991), "Credit Rationing and the Demand for Owner-Occupied Housing," *Journal of Urban Economics* 30 (July): 48–63.

Schreft, Stacey L., and Raymond E. Owens (1991), "Survey Evidence of Tighter Credit Conditions: What Does It Mean?" Federal Reserve Bank of Richmond *Economic Review,* March/April, 29–34.

Stiglitz, Joseph E., and Andrew Weiss (1981), "Credit Rationing in Markets with Imperfect Information," *American Economic Review* 71 (June): 393–410.

Townsend, Robert M. (1979), "Optimal Contracts and Competitive Markets with Costly State Verification," *Journal of Economic Theory* 21 (October): 265–93.

Williamson, Stephen D. (1986), "Costly Monitoring, Financial Intermediation, and Equilibrium Credit Rationing," *Journal of Monetary Economics* 18 (September): 159–79.

Measuring the Benefits of Unilateral Trade Liberalization Part 1: Static Models

Carlos E. J. M. Zarazaga

The ultimate object of interest is not the outcomes the model produces but how those outcomes affect society's welfare.

Advocates of free trade in the Americas may have been disappointed last year when the U.S. Congress denied the president fast-track authority to pursue trade agreements with other nations. The main concern of free trade advocates is that this decision will halt, and even reverse, the trend toward free trade initiated in the Americas in the 1980s. This frustration would be particularly justified if countries in the region would benefit from mutual trade concessions but not from a unilateral move to free trade.

Should countries pursue free trade policies even when their trading partners do not? Readers familiar with Ricardo's celebrated comparative advantage theory of international trade would probably say yes.[1] However, contemporary models of international trade do not warrant such a clear-cut conclusion. This is the first of two articles that examine the reasons behind this ambiguity.

This first article examines the welfare gains from unilateral trade liberalization suggested by static models of international trade. The second article will study the welfare gains from unilateral trade liberalization predicted by the more realistic (but also more complicated) dynamic models.[2]

A survey of the literature reveals that models predicting that unilateral free trade will be harmful to societies rely on a common assumption whose empirical and theoretical foundations have not been convincingly substantiated. Thus, I conclude that, on balance, the arguments in favor of a unilateral move to free trade are stronger than those against it.

QUANTIFYING THE BENEFITS FROM FREE TRADE

Researchers trying to quantify the effect of tariff changes on international trade do so in what seems a logical way: they choose a state-of-the-art theoretical model deemed appropriate for the case under study, assign values to the model's parameters and variables, and measure its quantitative performance under various trade policies.

This was essentially the strategy used to evaluate the benefits of the North American Free Trade Agreement (NAFTA) for all countries involved. The so-called general equilibrium models of trade were the state of the art at that time. The qualifier "general" differentiates these models from a previous vintage of "partial" equilibrium models, which assume prices in some markets are given and somehow determined outside the model.

For example, in analyzing the effects of tariffs, many partial equilibrium models assume

Carlos E. J. M. Zarazaga is a senior economist and executive director of the Center for Latin American Economics of the Federal Reserve Bank of Dallas.

real wages are fixed. This is highly unrealistic because tariffs usually alter the demand for factors of production by the industry or sector being protected, which in turn is likely to affect the relative prices of labor and/or capital. General equilibrium models, instead, allow for these effects because prices and quantities are determined endogenously, that is, within the model.

Theoretical General Equilibrium Models

The first element necessary in evaluating the benefits from unilateral trade liberalization is an appropriate general equilibrium model. General equilibrium models attempt to mimic as closely as possible actual economies by constructing an artificial (or model) economy the researcher can experiment with on the computer. This artificial economy is an abstract mathematical representation of the environment in which relevant economic agents are thought to operate and of the decision process by which those agents are thought to make their choices of consumption of different goods, of accumulation of capital, and so on. Methodologically, this implies that any general equilibrium model must start by specifying endowments, preferences, and technology.

The specification of preferences is an important step in formulating a general equilibrium model because the ultimate object of interest is not the outcomes the model produces but how those outcomes affect society's welfare.

For example, a reduction of tariffs on capital goods from *x* percent to *y* percent may double the rate of investment. The resulting increase in the capital stock will bring about higher growth, making it tempting to conclude that this higher growth will benefit society. However, accumulation of capital requires saving, which necessarily takes place at the expense of current consumption. If it were true that societies are always better off with faster growth, governments throughout the world could readily gather the necessary political support to adopt draconian measures reducing consumption by up to 50 percent. As history has proved, such a Stalinist approach to growth is doomed because it will be resisted by current generations, the ones who have to pay for that future growth with a drastic reduction in their consumption.

It is not a given, therefore, that the benefits of a reform (be it a trade reform or a tax reform) should be measured by the additional growth eventually made possible by the reform. Clearly, it would be preferable to measure those benefits in terms of the additional well-being the reform brings about. Such a task requires the construction of an index of well-being, or welfare, with some ideal properties. Unfortunately, construction of such an ideal index is impossible, as Nobel Prize recipient Kenneth J. Arrow (1951) demonstrates in his celebrated impossibility theorem. Yet economists are asked to evaluate the costs and benefits of a reform and, therefore, must provide a connection between observable outcomes (such as investment, growth, or consumption) and some measure of well-being, however imperfect or debatable.

General equilibrium models assume the welfare of a typical household provides a good approximation to social welfare. Such an approach implicitly assumes all households and consumers have well-defined preferences over different economic outcomes and that such preferences have a mathematical representation; that is, they can be measured by some utility function. Usually, this utility function is assumed to depend on the quantity of goods and services households consume. The functional form assumed for the utility functions is not completely arbitrary. It is in part dictated by restrictions imposed by basic axioms of consumer theory. One such restriction is that consumers always prefer to have something of *every* good rather than a lot of some goods and nothing of others.[3] In addition, a standard assumption is that preferences and the utility function that represents them are identical across households. This assumption guarantees that the utility function of any household adequately summarizes the welfare of all households.

A general equilibrium model exploits these assumptions by proposing that the representative household of a given country derives welfare from consuming, for example, the only good produced and exported by the country (c_1) as well as from consuming the only good imported by the country (c_2) and that the welfare this representative consumer obtains from different bundles of these goods can be measured by a welfare or utility function such as

$$(1) \qquad \textit{Welfare of representative consumer} = \alpha_1 \log c_1 + \alpha_2 \log c_2,$$

where α_1 and α_2 are parameters that measure the relative importance the representative consumer attaches to each good in his preferences.

Given the focus of this article, it is worthwhile to note that the level of tariffs does not appear explicitly in Equation 1. Changes in tariffs, such as those during trade liberalizations,

appear only indirectly in the welfare function, to the extent that they induce changes in the outcomes (such as the consumption level) over which consumers define their preferences.[4]

Another important element in the abstract construct of general equilibrium models is the postulate that economic agents act purposefully to achieve the ends they seek. Consistent with this methodological approach, households are assumed to maximize their level of welfare, or utility (as measured by the utility function), subject to the limitation imposed by their income or budget constraint. The budget constraint is a mathematical representation of the commonsense principle that households cannot spend more than their revenues from all sources (capital and labor income, savings carried over from the past, credit).

The consumer's maximization problem described above is not trivial because consumers can purchase different consumption bundles with their available income, but not all those bundles deliver the same level of utility. The solution to the problem requires finding the consumption bundle that allows the consumer to achieve the maximum possible welfare at given prices. Solving the problem repeatedly for different prices usually delivers well-defined demand functions—the standard, textbook, downward-sloping demand curve for each good. (For a more formal presentation, see the box titled "The Decision Problem of Consumers and Firms in General Equilibrium Models.")

Notice that changes in tariffs will generally change the prices of the goods from which consumers obtain utility. The price changes will alter the budget constraint, which may in turn change not only the consumption bundle that maximizes welfare but the level of welfare itself. That eventual shift in welfare induced by changes in tariffs is what general equilibrium models seek to measure.

A partial equilibrium model would stop the analysis here, in what could be referred to as the consumers', or demand, side of the economy. However, the endogenous determination of equilibrium prices, which result from the interaction of supply and demand, is in the very nature of general equilibrium models. Therefore, general equilibrium models must specify the suppliers', or production, side of the economy as well.

To that end, firms are assumed to combine primary factors of production (labor and capital) to maximize their profits. The transformation of these factors into output takes place according to some technology, mathematically represented by a production function.

Unfortunately, an issue of controversy among economists is whether production functions are characterized by constant returns to scale or increasing returns to scale. A 20 percent increase of all inputs results in a 20 percent increase in output under constant returns to scale but in a, say, 30 percent expansion of output under increasing returns to scale. The controversy is relevant to welfare gains from free trade because such gains tend to be larger under increasing returns to scale (see the box titled "The Decision Problem of Consumers and Firms in General Equilibrium Models").

General equilibrium models connect the household and firm sectors of the economy by exploiting the fact that households are the ultimate owners of the factors of production—labor and capital—and, therefore, the ultimate recipients of the factor payments and profits the firms make.

Finally, the international link in general equilibrium models of international trade is provided by assuming each country exports those goods for which domestic output exceeds domestic consumption. The model is one of general equilibrium in the sense explained earlier, that prices are set endogenously at the level necessary to ensure the quantities supplied and demanded of all goods, services, and factors are equal.

Applied General Equilibrium Models

Applied general equilibrium models attempt to exploit the theoretical framework offered by general equilibrium models to answer specific quantitative questions such as what the welfare gains are from a particular trade agreement (such as NAFTA) or from unilateral trade liberalizations. A theoretical general equilibrium model is brought down to earth by assigning concrete values, for example, to the parameters α_1 and α_2 in Equation 1.

Another difference between theoretical general equilibrium models and applied ones is in the number of economic sectors they can handle. Theoretical models are concerned with analytical and general results, which are almost impossible to derive in a large model. By contrast, applied models are more interested in quantitative answers to specific problems and situations. Free from the obligation to deliver general results and theorems, applied general equilibrium models can specify a large number of economic sectors, as many as necessary to accomplish the desired level of realism, the only limitation being the computational ability to solve the model numerically.[5]

The Decision Problem of Consumers and Firms in General Equilibrium Models

Any general equilibrium model that attempts to measure the impact of trade policies on welfare must start by postulating the utility, or welfare, function of the representative household populating the artificial, or model, economy.

For example, Harris (1984) proposed to evaluate the welfare effects of trade liberalization according to the following utility, or social welfare, function:

(B.1) $$Welfare = \sum_i \alpha_i \log c_i,$$

where c_i is the real consumption of good i and the summation is over all goods i, and α_i is a parameter that measures the importance of good i in households' preferences.

General equilibrium models postulate that economic agents act purposefully to achieve the ends they seek. Hence, households are assumed to maximize their level of welfare, or utility (as measured by the utility function), subject to a budget constraint.

In Harris' study, the representative household is endowed with an exogenous real income I. The consumer's problem can be represented, in the abstraction of a general equilibrium model, as the problem of maximizing Equation B.1 subject to the budget constraint

$$\sum_i p_i c_i \leq I,$$

which says that the sum of the price (in real terms) times the quantity purchased of each good (total expenditures) over all goods should not exceed total real income I.[1]

Several consumption bundles will satisfy the budget constraint above, but only one will maximize the preferences, or welfare, given by Equation B.1. The consumer's decision problem consists of finding such a bundle. In this example, a standard first-order-conditions approach delivers the answer mathematically.

Notice that removal of international trade tariffs will generally change prices and thus the budget constraint and optimal consumption bundle. Through this channel the imposition or removal of tariffs affects welfare in general equilibrium models.

Recall that, in this type of model, prices are not taken as coming from outside the model but rather determined inside the model from the interaction of supply and demand. The demand side of the economy is characterized by the consumer's maximization problem just described, while the supply, or production, side is characterized by the firm's maximization problem described below.

General equilibrium models assume that firms combine primary factors of production (labor and capital) to maximize their profits. The transformation of capital and labor into output takes place according to some technology, mathematically represented by a production function.

To this end, the profits of a typical firm (or industry) i are represented mathematically as

$$Profits = p_i y_i - wl_i - rk_i,$$

where $p_i y_i$ represents the firm's revenues (from selling a quantity y_i of good i at real price p_i) and the second and third terms are the costs of producing output y_i, where wl_i represents the labor costs associated with hiring l_i hours of labor at the hourly real wage w and rk_i the cost of renting k_i units of capital at the rental price r.[2]

The technological constraint, in turn, is represented as

$$Output\ of\ firm\ i = y_i = F(l_i, k_i),$$

where F is some function of l_i and k_i. As explained in the text, the nature of that function is a controversial issue. In particular, there is disagreement as to whether the production function F of the typical firm is characterized by constant returns to scale or increasing returns to scale.

For example, a possible mathematical representation of the production function F is

$$Output\ of\ firm\ i = F(l_i, k_i) = A * l_i^{\gamma_1} * k_i^{\gamma_2},$$

where $A > 0$, k_i is the level of capital, l_i is the amount of labor input and $\gamma_1 > 0$ and $\gamma_2 > 0$ are parameters. When $\gamma_1 + \gamma_2 = 1$, the above production function is constant returns to scale, as the reader can verify by multiplying capital and labor by the same percentage increase x. In that case:

$$A * \left[\left(1+\frac{x}{100}\right) * l_i\right]^{\gamma_1} * \left[\left(1+\frac{x}{100}\right) * k_i\right]^{\gamma_2}$$

$$= \left(1+\frac{x}{100}\right)^{\gamma_1} * \left(1+\frac{x}{100}\right)^{\gamma_2} * A * l_i^{\gamma_1} * k_i^{\gamma_2}$$

$$= \left(1+\frac{x}{100}\right)^{\gamma_1+\gamma_2} * A * l_i^{\gamma_1} * k_i^{\gamma_2} = \left(1+\frac{x}{100}\right) * A * l_i^{\gamma_1} * k_i^{\gamma_2},$$

which says increasing each input by x percent results in an increased output of also x percent.

Note, however, that with $\gamma_1 + \gamma_2 > 1$, this same production function becomes increasing returns to scale: an increase of x percent in each input results in a larger proportional increase of output—specifically, in an increase of

$$\left(1+\frac{x}{100}\right)^{(\gamma_1+\gamma_2)} > 1+\frac{x}{100}.$$

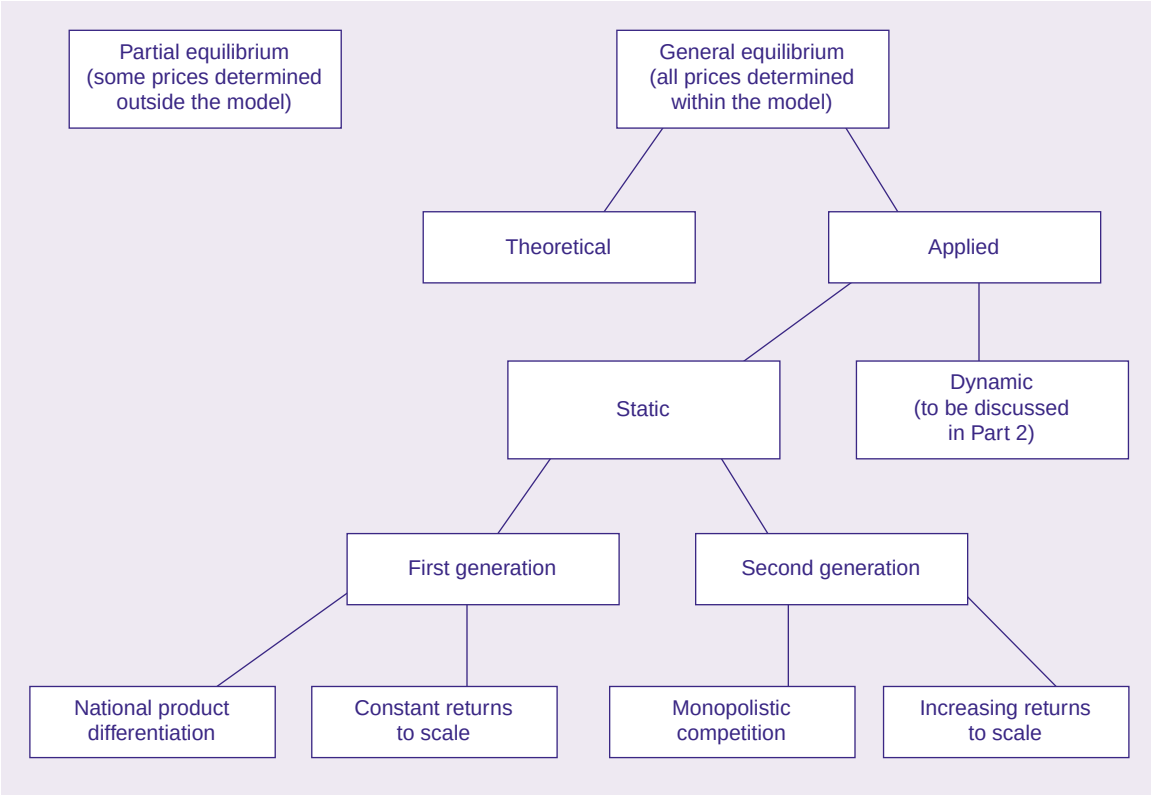
Unfortunately, theoretical considerations do not permit exclusion of either case, and the empirical evidence is mixed. This is somewhat problematic because, as stated in the text, gains from free trade tend to be larger under increasing than under constant returns to scale.

Finally, note that in general equilibrium models the household and firm sectors are connected because households' real income I (the right-hand side of their budget constraint) is nothing but the sum of the firms' profits and their payments to labor and capital inputs.

[1] Implicitly, all nominal quantities (prices and income) are being deflated by a common price index P.

[2] The rental price of capital, r, is generally given by the sum of the real interest rate and the depreciation rate.

Figure 1
Models of International Trade



Once the researcher has defined the utility function, the production functions, the number of sectors to be considered, and concrete values for the relevant parameters, it is possible to quantify the impact of a policy change by computing the model for different trade policies. The different policies' impact on welfare is then analyzed by reporting, for example, that in the artificial economy the welfare gains after the policy change are such that GDP should be *x* percent higher to achieve that same welfare without a reform. This measure is often referred to as the equivalent variation in income.

Different applied general equilibrium models give different qualitative answers to the question of whether unilateral free trade is a wise policy. These differences extend to the quantitative importance of the welfare gains or losses from such a move. The remainder of this article examines the details of the models responsible for those discrepancies.

STATIC APPLIED GENERAL EQUILIBRIUM MODELS OF FREE TRADE

Static applied general equilibrium models ignore the time dimension. The artificial economies of these models have no past or present and, therefore, no incentive to save or lend. As a consequence, countries cannot run a current account balance or trade deficit.

For purposes of exposition, and following the classification Brown (1992) proposed, the static applied general equilibrium models of unilateral trade liberalization can be grouped into two categories: first- and second-generation (*Figure 1*). First-generation models have in common two assumptions: the national product differentiation assumption and the constant-returns-to-scale assumption. The second-generation models replace the national product differentiation assumption with the monopolistic competition assumption, and the constant-returns-to-scale assumption with the increasing-returns-to-scale assumption. The remainder of this section analyzes how those assumptions affect the different welfare results from unilateral trade liberalization delivered by those models.

The Role of the National Product Differentiation and Constant-Returns-to-Scale Assumptions

Because their focus is empirical rather than analytical, applied general equilibrium models must be able to interpret actual data as

reported in official or private statistics. Unfortunately, available data are often not the exact empirical counterpart of a concept or definition used in the theoretical model. This discrepancy may force a compromise between theory and reality that can weaken confidence in the empirical results of a model.

In the case of international trade, researchers formulating theoretical models would like to make the sensible assumption that consumers don't discriminate goods by their origin. In other words, consumers value a good supplied by one country as much as the same good produced by another country. The assumption that a good is a perfect substitute in demand across origins seems natural in any model dealing with international trade. One empirical implication of this perfect-substitution assumption is that countries will import goods different from those they export, because countries typically export their production surplus after satisfying domestic demand for the good.

Under this perfect-substitution assumption we shouldn't see, as we do, trade statistics reporting that Germany and Japan import *and* export cars. Actually, the cars Germany imports differ from those it exports. Germany may export BMWs to Japan and import Hondas or Toyotas from Japan. Less obvious distinctions can be made as well. For example, a country may export two-door cars and import four-door cars; it may import cars with sunroofs and export cars without them, and so on. However, because such details are lost in trade statistics, it may appear as if countries import and export the same kind of goods. This seemingly puzzling situation, known among international trade scholars as the cross-hauling problem, occurs not because the perfect-substitution assumption is unrealistic but because of the way trade data are recorded. There are hundreds, perhaps thousands, of varieties of cars, and trade statistics should report, strictly speaking, exports and imports for each of them. But processing the information in such detail would be costly; in practice, many kinds of cars are grouped under broad categories. This renders impractical theoretical models based on the perfect-substitution assumption.

On the other hand, the nature of the problem suggests the way around it. If trade statistics fail to recognize that a good being imported is not actually the same as the good being exported, the solution is to assume that goods reported as both exported and imported are really different. In other words, the solution is to assume goods differ not only by type but also by origin. This means, following our example, that cars produced abroad are not the same as cars produced domestically, which is a rather ingenious way to distinguish the BMWs Germany exports from the Toyotas Germany imports. This is the national product differentiation assumption adopted by the first generation of applied general equilibrium models.[6]

The national product differentiation assumption implies that the goods produced by each country are unique and, therefore, cannot be perfectly substituted by any of the goods produced by any other country. Although this assumption solves the problem of the lack of correspondence between data and theory otherwise present with the assumption of perfect substitution of goods across origins, it introduces a new problem: now countries have monopoly power over the goods they produce. The reason, of course, is that no other country can produce the same good. In the logic of this assumption, in a world where there are n products and m countries, there will be $n*m$ goods.

Most first-generation applied general equilibrium models of international trade combine the national product differentiation assumption with a constant-returns-to-scale technology. The combination of these two assumptions has serious theoretical and quantitative implications for the analysis of trade liberalization because, in the presence of market power, only imposing tariffs—not reducing them—improves a country's welfare.

The assumption of constant returns to scale is important because it keeps the market power at the country rather than at the firm level. Under constant returns to scale, the marginal cost is constant, independent of the level of production. Therefore, all firms will supply their outputs at a price equal to the constant marginal cost. Any attempt by an individual firm to set a higher price will divert its customers to competitors. Of course, no firm will set prices below marginal cost because it would be producing at a loss. None of the firms can, individually, exploit the market power implicit in the fact that no other country can produce the same products they do. Thus, the government can intervene by coordinating the firms' actions to enable them to exploit their market power. For instance, the imposition of a tax (tariff) on foreign goods will increase the domestic price of imports relative to the domestically produced—and eventually also exported—good. The lower relative price of the domestic good will induce more consumption of it and less of the imported good, producing two effects. On the one hand,

it will reduce the demand for the foreign good by the tariff-imposing country and, therefore, generate downward pressure on the world price of that good. On the other hand, it will reduce the surplus of the domestic good available for export to world markets, which will increase the international price of the good. This implies that the terms of trade—that is, the international price of exports relative to that of imports—shift in favor of the tariff-imposing country. (For a more detailed explanation of this result, see the box titled "Optimal Tariff Under the National Product Differentiation Assumption.")

Reversing the argument, the unilateral removal of a tariff can worsen the terms of trade and be welfare-reducing, especially if the tariff had been at the level at which a country exploits its market power the most.

Welfare Gains from Unilateral Trade Liberalization in First-Generation Applied General Equilibrium Models

The few static applied general equilibrium models that have attempted to measure the gains of unilateral trade liberalization for a small country have indeed found negligible, or even negative, welfare gains from a unilateral move to free trade. Boadway and Treddenick (1978) found that removal of tariffs in Canada would cause welfare to decline by about 1 percent or increase by only 0.06 percent. The terms-of-trade deterioration resulting from an import tariff reduction, as implied by the national product differentiation assumption, has led Brown (1987) to conclude rather categorically that unilateral trade liberalization is rarely welfare-improving, even for a small country, in this first generation of static applied general equilibrium models.

It is important to remember, in evaluating those disappointing welfare results for the cause of free trade, that the motivation for the national product differentiation in first-generation applied general equilibrium models was mainly pragmatic, an apparently innocuous way to bridge the gap between theory and available data.[7] However, this compromise may not appear as appealing when it becomes apparent that the assumption, in combination with constant returns to scale, implies that a country (not its industries) has complete monopoly power in the market for its exports and that this market power introduces a bias against trade liberalization. This bias would not be problematic if the market power implication of the national product differentiation assumption could be empirically validated, but this may not always be the case. These considerations led to the formulation of second-generation static applied general equilibrium models, which replace the assumption of national product differentiation with the monopolistic competition assumption. The next section explains why second-generation models deliver somewhat higher welfare gains from unilateral trade liberalization.

The Role of Monopolistic Competition and Increasing Returns to Scale

To correct the country-monopoly-power side effect introduced by the national product differentiation assumption, many authors have replaced it with the assumption that each firm, rather than each country, produces a different product, transferring the monopoly power from the country to the firm level. This monopoly power is limited, however, by the fact that consumers can easily substitute the products of one firm with close varieties of the same good produced by another firm. Technically, each firm produces an imperfect-substitute good under monopolistic competition conditions.

Because each firm specializes in the production of a good no other firm can produce, the firm is able to exploit its market power on its own, without the help of an import tariff levied by the government. That is, firms exploit their market power as much as they can before any tariff is imposed. Consequently, the imposition of a tariff under monopolistic competition will be not only redundant but also, in general, detrimental to society. Not surprisingly, applied general equilibrium models relying on monopolistic competition will tend to find that unilateral removal of tariffs is welfare-improving.

The introduction of monopolistic competition in applied general equilibrium models solves the same problem the national product differentiation assumption does and at the same time avoids this assumption bias against unilateral trade liberalization. In particular, monopolistic competition can still account for the considerable cross-hauling observed in trade statistics. The puzzling observation that a country appears to export the same product it imports can be interpreted as a domestic firm producing (and exporting) a variety of the product different from the one being imported.

Unfortunately, the monopolistic competition assumption has a drawback the national product differentiation assumption does not have: it implies that each product variety will be produced by one and just one firm. This implication is problematic because it conflicts with the standard assumption of constant-returns-to-scale production technology.

Recall that a constant-returns-to-scale technology can deliver a given percentage change in the output of a good by simply changing all the inputs by that same percentage. This means any level of output Q_N of a certain good can be produced either by a single firm or by any number N of identical firms, each of them using $1/N$ fewer inputs than a single firm would to produce Q_N. In other words, under constant returns to scale the output of any firm can be replicated by N smaller and alike firms, yet this technologically natural possibility would be ruled out by the monopolistic competition assumption that each product can be produced by only one firm.

To save the contradiction of simultaneously assuming product differentiation and a constant-returns-to-scale technology at the firm level, most trade models appealing to monopolistic competition also assume the technology is increasing returns to scale. Equivalently, second-generation applied general equilibrium models of international trade assume the total production cost is composed of two parts: a fixed cost independent of the level of production and a variable cost proportional to the level of output.

In its simplest form, this assumption takes the mathematical representation

$$Total\ cost = F + bQ,$$

where F is the fixed cost and b the cost of an additional unit of output Q—that is, the marginal cost. Krugman (1979) uses this formulation to show how increasing returns to scale can account for international trade. What is important about this technological specification is that the *average* cost declines with the level of output. This can be seen easily by dividing the above equation by the level of output

$$Average\ cost = TC/Q = F/Q + b.$$

Since F is a fixed number, the ratio F/Q declines as Q, the level of output, increases, thus reducing overall average cost. This is exactly what one would expect from a technology under which a given percentage increase in inputs (and, therefore, in costs) delivers an even higher proportional increase in output.

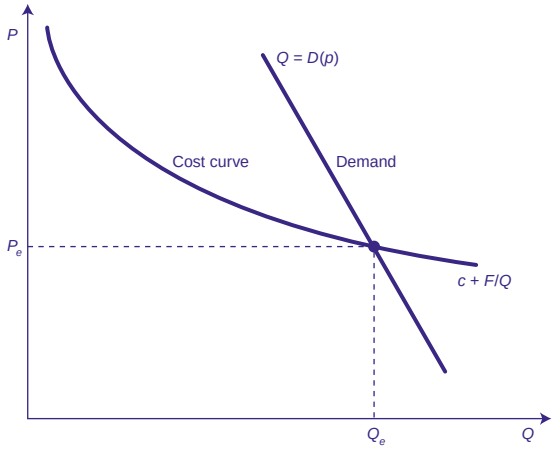
The shape of the average cost curve of a hypothetical firm with an increasing-returns-to-scale technology is depicted in Figure 2, which also displays the hypothetical demand curve for the good produced by the hypothetical firm. The figure suggests one of the main implications of the increasing-returns-to-scale assumption: only one firm will produce each good.

To see why, suppose the hypothetical firm of Figure 2 produces all output of a given good. The market equilibrium for that good will occur at the price P_e, where the firm will be able to satisfy the quantity demanded Q_e and at the same time cover the costs of doing so, since the price P_e equals the average cost at that level of output.

Now suppose N firms were going to supply the market. As in the case of constant returns to scale, each of them will operate at a lower scale than a single firm. But unlike in the constant-returns-to-scale case, the average cost for each of the firms will be higher than it would be for a single firm. This implies they would have to charge a higher price than the single firm would, that is, a price higher than P_e. Realizing this, at least one of these firms sooner or later will try to capture competitors' customers by cutting the price to P_e. Such a firm will be able to sustain this lower price without losing money because it will now supply the whole market, and under increasing returns to scale it will be able to produce that larger quantity at the lower average cost P_e.[8]

The prediction that just one firm will produce each good under increasing returns to scale is logically consistent with the monopolistic competition assumption that each good will be produced by only one firm. For this reason, second-generation applied general equilibrium models adopted the increasing-returns-to-scale assumption along with the monopolistic competition approach.

Figure 2
Increasing Returns to Scale and the Number of Firms



Optimal Tariff Under the National Product Differentiation Assumption

This box explains in more detail why the optimal tariff tends to be strictly positive in models of international trade that use the national production differentiation assumption when the production function at the firm level is assumed to be constant returns to scale.

To keep the discussion as nontechnical as possible, it will be presented in terms of standard graphical representations of the welfare function and budget constraint, although a more rigorous mathematical representation is possible and available in many advanced international trade textbooks.[1]

Figure B.1 presents the standard two-dimensional representation of a welfare function such as that in Equation 1 in the text. Each of the curves in the figure traces the combinations of the quantities consumed of good 1 (c_1) and of good 2 (c_2) that allow the typical consumer of a given country (let's say the home country) to attain the same level of utility or welfare. For example, the curve labeled $U = 4.61$ represents the different pairs (c_1, c_2) from which the consumer can derive a utility level of 4.61 when preferences are represented mathematically by Equation 1, with parameter values $\alpha_1 = \alpha_2 = 0.5$. The reader can verify, using Equation 1, that the same level of welfare can be attained with the consumption pairs $c_1 = 50$ and $c_2 = 200$, $c_1 = 100$ and $c_2 = 100$, or $c_1 = 200$ and $c_2 = 50$.

The straight line running from A to B in Figure B.1 is a geometric representation of the typical consumer's static budget constraint of the home country for the case of two goods. This line represents the different quantities c_1 and c_2 of each good the home country consumer can afford to buy when his income (in terms of good 1) equals A and the relative price of good 1 in terms of good 2 is 0.5. In what follows, and in line with the national product differentiation assumption, it will be assumed that the home country is the only world producer of good 1 and the foreign country is the only producer of good 2.

Now suppose the typical consumer of the home country is endowed with A units of good 1 and none of good 2. Because preferences are concave, meaning consumers have a taste for variety, the typical home country consumer would like some quantity of good 2 as well. However, according to the national product differentiation assumption, the home country is unable to produce good 2, so it is willing to trade part of its endowment of good 1 in exchange for some amount of good 2 produced by the foreign country. When the relative price, or terms of trade, of the home good (export) in terms of the foreign good (import) is 0.5 as assumed, trade between the home and foreign country will take place at the ratio of one unit of good 1 for half a unit of good 2.[2] These particular terms of trade are represented in Figure B.1 by the slope of the budget constraint. This can be verified in the figure: when the typical home country consumer gives up consumption of 100 units of good 1, represented by the movement from A to C, the home country can export those 100 units to get in exchange 50 units of good 2, represented by the vertical distance $D - C$.

Figure B.1
Utility Maximization

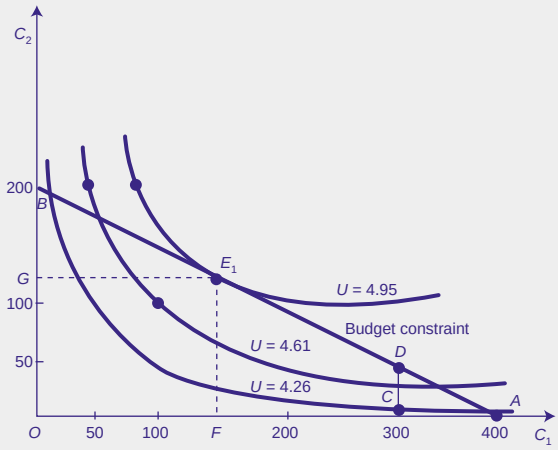
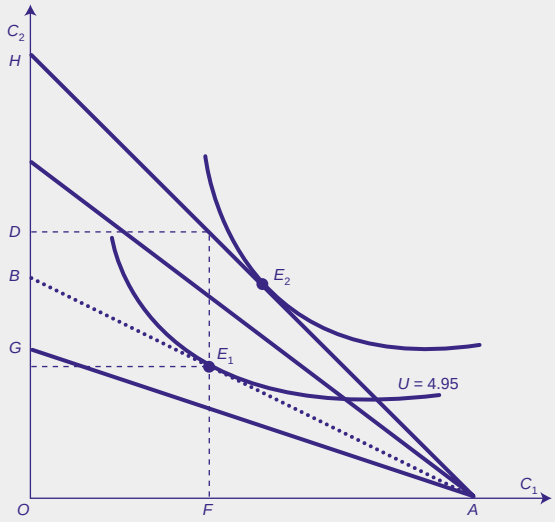


Figure B.2
Changes in Terms of Trade and Welfare



The home country consumers will keep trading good 1 for good 2 until they maximize their satisfaction or welfare, which will occur at point E_1 in the figure. Of all the combinations of c_1 and c_2 the consumer can afford with the budget AB, the one at point E_1 yields the highest level of welfare. This implies that the home country will consume the quantity F of good 1, export the quantity $A - F$ of that good—the excess of endowment of good 1 over the domestic consumption of good 1—and import the quantity G of good 2.

In the above example, the terms of trade were arbitrarily set at 0.5. But of course an entirely analogous analysis applies to any terms of trade. Figure B.2 represents a set of indifference curves along with several budget lines, each with a different slope and corresponding to different terms of trade. The dotted line represents the baseline case of the previous example in which the terms of trade were assumed to be 0.5. Budget lines above the dotted line represent improvements in the terms of trade for the home country with respect to the baseline case. For example, the budget line running from A to H has a slope of 1, which is greater than 0.5. At that budget line, the home country can trade each unit of the good it produces for one unit of the good it does not. This means the home country good (good 1) is now more valuable than before: it is worth as much as the imported good instead of only half, as it was when the terms of trade were 0.5. Of course, the home country consumers will benefit from the fact that the home good is more valuable relative to the foreign good. This improvement is reflected in that, with the new budget constraint AH, the optimal consumption of the typical home country household occurs at point E_2,

Optimal Tariff Under the National Product Differentiation Assumption (continued)

with a higher level of welfare than at E_1, the point representing the welfare-maximizing consumption pair when the terms of trade were 0.5.

The above analysis suggests that a country can improve its situation if it can influence the terms of trade in its favor. The imposition of a tariff under the national product differentiation assumption can do just that. Under this assumption each country has market power over the goods it produces because no other country produces those goods.

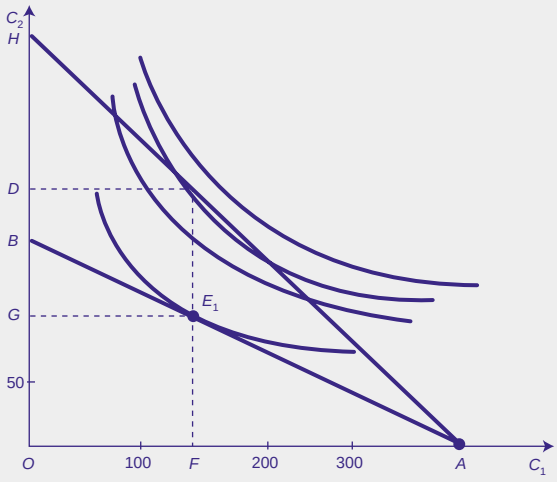
To see how a tariff can improve a country's situation, let's consider the baseline example when the terms of trade were 0.5. Now suppose the home country imposes a tariff of 50 percent on the imported good (good 2). For practical purposes, the imposition of a tariff implies that a certain amount of the imported good will have to be surrendered to customs officials. In this case, a tariff of 50 percent means that now each consumer will get 0.25 units of good 2 for each exported unit of good 1. The government will get to keep the remaining 0.25 unit. However, this analysis is valid only if the terms of trade remain the same after the imposition of the tariff, which is unlikely because the tariff makes the imported good more expensive relative to the home good for the home country consumers. They will consume less of the imported good and more of the exported good. The lower home country consumption of the imported good will put downward pressure on the world price of imports. Meanwhile, the higher domestic consumption of the home good will reduce the surplus of it available for export and put upward pressure on its international price. A fall in import prices and a rise in export prices will improve the terms of trade for the tariff-imposing country.

Suppose then that the imposition of a 50 percent tariff on good 2 by the country producing good 1 results in a 100 percent improvement in its terms of trade. In other words, suppose the terms of trade move from 0.5 to 1 as a result of the tariff. The new budget constraint implied by these new terms of trade for the home country consumers is represented in Figure B.3 by an outward movement of the budget line from *AB* to *AH*. At these new prices, the typical household of the home country will export $A - F$ units of good 1, receive in exchange D units of good 2, surrender half this amount, $D - G$, to customs, and keep the remaining half, G, for its own use. From the home country consumer's perspective, nothing has changed. He still receives half a unit of good 2 for each unit of good 1 he exports (or, equivalently, for each unit of good 1 he does not consume). The optimal consumption combination will still be represented by point E_1, where the typical household attains the same level of welfare as before the imposition of the tariff.

It would appear, then, that imposing the tariff was a futile policy move. However, the country as a whole is richer because now the government has extra revenues, represented by import duties $D - G$ collected by customs. The government can use the additional income to upgrade public services. Or it can return the revenues to households in the form of higher pensions, social security benefits, or income tax reductions. The households, in turn, can use this tax rebate to increase purchases of both goods 1 and 2 and, therefore, to attain higher levels of utility than represented by point E_1. The home country's welfare is improved after the tariff because the government successfully transfers the burden of the tariff to foreigners by altering the terms of trade against them.[3]

Of course, this example was designed to deliver welfare gains from a tariff. It is also possible that the terms-of-trade gain from a 50 percent tariff is too small to compensate the home country for the distortionary costs from a tariff. The ultimate benefit of a tariff depends on, among other things, the elasticities of demand for the home and foreign products both at home and abroad. The details are beyond the scope of this article; however, the interested reader is referred to Brown (1987), in which it is also argued that, empirically, most models using the national product differentiation assumption produce terms-of-trade gains strong enough to improve the welfare of the tariff-imposing country.

Figure B.3
Welfare-improving Tariff



It may seem odd that a society can gain from imposing a tariff. After all, a tax on imports—like any other tax—is, in principle, welfare-decreasing. However, this outcome ignores the presence of an externality under the national product differentiation assumption: the country has a market power that its firms cannot exploit at an individual level when their production technology is constant returns to scale. This sort of externality justifies the government intervention. Under those circumstances, a tariff turns out to be corrective rather than distortive. By imposing a tariff, the government implicitly forces the home country firms and households to coordinate their actions to produce and consume, in the aggregate, the quantity of good 1 at which the home country exploits its market power the most.

The result that the optimal tariff is eventually positive will tend to disappear, therefore, in models that assume increasing returns to scale at the firm level or that abandon the national product differentiation assumption and its implicit market power. In the first case, as explained in the text, each firm will be able to exploit its market power on the differentiated good it produces, without the need for the government to step in. In the second case, many different countries will be producing each good; therefore, no individual country will have control over the market of any particular good. However, in face of the considerable amount of cross-hauling of goods across borders reported in international trade data, abandoning the national product differentiation assumption may be problematic for addressing some quantitative questions in international trade.

[1] See, for example, Grubel (1977), especially chap. 8, p. 155.
[2] These terms of trade imply that the good produced by the home country is worth, in dollar terms, half the price of good 2.
[3] Notice that now the foreign country receives less for its exports and pays more for its imports.

Welfare Gains from Unilateral Trade Liberalization in Second-Generation Applied General Equilibrium Models

As previously explained, compared with national product differentiation, monopolistic competition is more favorable in principle to the case of free trade because in such an environment firms incorporate their market power in their price decisions, so no corrective government tariff is needed.

This intuition is rigorously confirmed by Harris (1984), who finds that welfare gains from unilateral trade liberalization are nil in the models with just national product differentiation and constant returns to scale, whereas they are in the range of 2.7 percent to 4.1 percent of GDP when some of the goods are produced under conditions of monopolistic competition and increasing returns to scale.

Although the gains from unilateral trade liberalization delivered by second-generation applied general equilibrium models are positive, they are far from staggering. After all, a gain in GDP of 3 percent to 4 percent is in the same order of magnitude as the normal annual GDP growth for many developed countries.

CONCLUSION

Static applied general equilibrium models fail to deliver the eye-popping gains from unilateral trade liberalization that free trade advocates often promise. As this conclusion applies to both first- and second-generation models, it appears to be rather robust to the significantly different assumptions about market structure or technology made by the two generations of static models.

However, static models are single-period models unable to incorporate the important dimension of time, ruling out savings and investment. Why should economic agents save and invest if there is no tomorrow? This is unfortunate in the context of trade liberalization because reductions of tariffs on, for example, capital goods, may induce more investment, which will increase capital stock and, hence, a society's ability to produce and consume more in the future. But static models—having neither past nor future—will leave out of the equation those welfare gains from larger future consumption, which can be far more important than any gains from larger present consumption. In fact, the omission of the time dimension may ultimately explain the negligible welfare gains from unilateral tariff reductions delivered by the static models examined in this first article. This is precisely the conjecture that will be more fully explored in the second article on this topic.

NOTES

[1] David Ricardo (1817) was the first economist to argue that free international trade would be beneficial to two countries even if one of them produced all traded goods more efficiently than the other. His key insight was that what makes international trade desirable is not this absolute advantage but the comparative or relative advantage—that is, the efficiency of each country at producing one good relative to its efficiency at producing other goods.

[2] Static models are those in which the dimension of time is missing. In such models there is no past or future: all analysis is conducted as if everything happened at one time. In contrast, dynamic models do incorporate the dimension of time in the analysis.

[3] In economists' technical jargon, preferences are concave.

[4] Notice also that the utility function given in Equation 1 satisfies the taste for variety mentioned in the text, in the sense that the consumer is always better off consuming a little of every good than a lot of one good and nothing of others. Demonstrated mathematically, the consumer would get an infinitely negative utility by consuming nothing of some good because in that case $\log 0 = -\infty$.

[5] This is why applied general equilibrium models are also referred to in the literature as computable general equilibrium models.

[6] Because Armington (1969) was the first to propose this assumption, it is often referred to in the literature as the Armington assumption.

[7] Another reason for the introduction of the national product differentiation assumption is that slight changes in tariffs tend to produce unrealistically large moves toward specialization in models making the opposite assumption that products are homogeneous across countries (see Shoven and Whalley, 1984, p. 1035).

[8] Readers familiar with the literature will recognize this argument as basically the contestable markets hypothesis often posed to question the effective market power of potential monopolies and their ability to depart too much from competitive outcomes.

REFERENCES

Armington, Paul S. (1969), "A Theory of Demand for Products Distinguished by Place of Production," *International Monetary Fund Staff Papers* 16 (March): 159–78.

Arrow, Kenneth J. (1951), "Social Choice and Individual Values," *Cowles Commission Monograph No. 12* (New York: Wiley).

Boadway, Robin, and John Treddenick (1978), "A General Equilibrium Computation of the Effects of the Canadian Tariff Structure," *Canadian Journal of Economics* 11 (August): 424–46.

Brown, Drusilla K. (1987), "Tariffs, the Terms of Trade, and National Product Differentiation," *Journal of Policy Modeling* 9 (Fall): 503–26.

——— (1992), "The Impact of a North American Free Trade Area: Applied General Equilibrium Models," in *North American Free Trade: Assessing the Impact,* ed. Nora Lustig, Barry P. Bosworth, and Robert Z. Lawrence (Washington, D.C.: Brookings Institution), 26–68.

Grubel, Herbert G. (1977), *International Economics* (Homewood, Ill.: Richard D. Irwin).

Harris, Richard (1984), "Applied General Equilibrium Analysis of Small Open Economies with Scale Economies and Imperfect Competition," *American Economic Review* 74 (December): 1016–32.

Krugman, Paul R. (1979), "Increasing Returns, Monopolistic Competition, and International Trade," *Journal of International Economics* 9 (November): 469–79.

Ricardo, David (1817), *The Principles of Political Economy and Taxation*, Reprint, London: Dent, 1962.

Shoven, John B., and John Whalley (1984), "Applied General Equilibrium Models of Taxation and International Trade: An Introduction and Survey," *Journal of Economic Literature* 22 (September): 1007–51.

Monetary Policy Arithmetic: Some Recent Contributions

Joydeep Bhattacharya and Joseph H. Haslag

This article explores some of the recent contributions to the literature on deficit financing and the unpleasant monetarist arithmetic.

Joydeep Bhattacharya is an assistant professor at Iowa State University. Joseph Haslag is a senior economist and policy advisor in the Research Department at the Federal Reserve Bank of Dallas.

Standard undergraduate textbooks often cast monetary policy and fiscal policy as separable undertakings. Such a split does seems natural; after all, the players involved are different entities. In the United States, for instance, monetary policy decisions are made by the Federal Reserve, while fiscal policies come under the purview of the federal government. A direct consequence of this "split personality" view of policy action is that it gives monetary policy sole authority over short-term nominal interest rates and/or money growth rates, while fiscal policy gets the final say on tax rates and transfer payment schedules. Indeed, in the monetarist–Keynesian debate, this separatist tradition asks which class of policies is more effective at managing economic activity.[1]

Almost two decades ago, Sargent and Wallace (SW) burst the bubble on this dichotomy in a pathbreaking 1981 article, arguing that neither policy is conducted in a vacuum.[2] Although it may be appropriate to think of monetary and fiscal policy actions as separate ventures, it is important to understand that the two interact. According to SW, monetary–fiscal policy distinctions are at best arbitrary; monetary policy actions have repercussions for fiscal policy settings and vice versa. Because governments, like private citizens, face budget constraints, SW show that both monetary and fiscal actions interact in a single, unified government budget constraint. Actions taken by the government while wearing the fiscal policy hat, for instance, eventually affect the actions it takes while sporting the monetary policy cap.

A convenient way to understand these monetary–fiscal policy interactions is to think of the central bank and the treasury as engaged in a game of chicken, from which, at most, one winner can emerge.[3] SW consider a setting in which the fiscal wing of the government dominates[4] and focus on a policy in which the treasury finances an increase in government spending by selling interest-bearing debt to the public.[5] Suppose that current money growth, at least initially, is unaffected by this fiscal policy action. In time, the real interest obligations of the treasury would rise. It is even possible that the revenue from new bond sales would be insufficient to pay the outstanding interest on past bond sales. When this rollover option fails, the government can potentially avoid bankruptcy by printing money to pay off the deficit.[6] SW show that this increased interest expense forces the government to print money at a faster rate than would have been necessary had it chosen at the outset to finance the deficit by printing money.

As a consequence, financing the deficit with bonds could ultimately be more inflationary than financing it by printing money. Throughout this article, we refer to unpleasant monetarist arithmetic (UMA)—a term SW coined—as an outcome in which money growth must rise to finance a permanent increase in government debt.[7]

Three conditions are needed to obtain this spectacular result in the SW setup: (1) the central bank is subservient to the fiscal authority, (2) the real interest rate on government debt is higher than the economy's real rate of growth, and (3) the central bank is in a position to raise revenue by printing money. For the SW result (which is contrary to conventional wisdom) to have empirical bite, it is important that some real-world economies share the three features of the SW model economy. If the original Sargent–Wallace UMA result is to serve as a cautionary note for policymakers, all three conditions must hold in the real world.

This article explores some of the recent contributions to the literature on deficit financing and the unpleasant monetarist arithmetic. Although government surpluses—not deficits—are currently making headlines, we believe it is premature to pronounce deficit financing dead, just as it is premature to declare the business cycle dead. After all, government surpluses are neither permanent nor universal. Where this discussion of the UMA may be most illuminating is in expanding our understanding of the deficit financing issues some developing countries face. Government bond sales (to finance deficits) in Russia, and more recently in Brazil, have coincided with a faster rise in inflation than money growth could explain. For countries with surpluses, the UMA's predictions may say something about disinflation.

This article begins with a brief statement and derivation of the UMA result, then reviews the evidence used to refute the UMA predictions. In particular, we present evidence on the real interest rate for the United States and Canada to check whether condition 2 is satisfied. We examine whether, in fact, condition 2 is necessary for the UMA result by considering a case in which we increase the number of assets people can use to transfer income across time. Could it be that with an asset structure less restricted than SW's, condition 2 is no longer necessary? Finally, we extend the SW analysis to consider the deficit-financing consequences of other monetary policy tools. Since central banks have more than one way to raise revenue, does it matter for the UMA which way is chosen?

THE ECONOMY

The starting point for our analysis is a stripped-down description of the economy in which a government is operating. Time is broken into discrete periods and indexed by $t = 1,2,\ldots$. At each date $t = 1$, N_t young people are born. Population grows according to the rule $N_{t+1} = nN_t$, where $n > 1$ is the gross rate of population growth. Each person lives two periods and lifetimes of agents overlap, so that a young person lives at the same time as an old person. The latter dies at the end of the period; the former moves into old age, and a new generation of young people is born. One group (the "initial old") enters date $t = 1$ with only one period of life left.

This economy has a single perishable commodity. Each person receives an endowment of y units of this consumption good when young (the period in which they are born) and nothing when old (the second, and last, period of their life). In this setup, population and aggregate income grow at the same rate.[8]

People in this economy wish to consume something when they are young and something when they are old. Because young people receive some of the consumption good only when young, each forges a plan that will maximize well-being from consumption over the course of a lifetime. These plans will require that each person consume a part of the endowment in the first period of life. What happens to the remainder of the endowment? Since the good itself is nonstorable, each person needs to purchase stores of value, which can be used to finance old-age consumption.[9] Let c_1 (c_2) denote the quantity of goods consumed when the person is young (old). Note that the young pay the government a lump-sum tax of τ goods. Therefore, the division of the endowment by the young person can be represented as

$$y - \tau = c_1 + s,$$

where s is the remainder that is used to buy stores of value. Let r denote the gross real return on the stores of value, so that the product, rs, represents the total goods a young person can consume when old. This means

$$rs = c_2.$$

A typical young person has access to two stores of value, money and bonds. The real purchasing power of the money held, vm, will, of course, change as monetary policy settings change. In the product, vm, m represents the pieces of paper money each young person holds and v is the quantity of the endowment

that can be acquired with one unit of paper money.[10] Hence, vm is the value of paper money (per young person) measured in terms of the consumption good. The young person may also buy treasury bonds, giving the treasury b endowment units in return for R/n units of the consumption good when the bonds mature (when the person is old). For simplicity, we assume the person consumes a fixed fraction of y when young and chooses a portfolio of both money and treasury bonds.[11] Put differently, the combined savings, $s = b + vm$, is independent of R/n.

Assume that the government taxes people only when they are young, collecting τ units of the consumption good. The government costlessly transforms the tax collected into units of a government good, denoted by g_t, that are useless to people. If the government wishes to acquire more units of the consumption good, it can borrow, issuing riskless interest-bearing bonds, B, that are repaid one period later. Alternatively, the government could print fiat money, M.[12] For each good borrowed from young people at date $t-1$, the government pays R_t goods at date t. Thus, at any date t the government's budget constraint is

$$(1) \qquad N_t g_t + R_t B_{t-1} = N_t \tau_t + B_t + v_t(M_t - M_{t-1}).$$

Equation 1 captures the required balance between what the government spends and what it collects. The total number of goods the government purchases is $N_t g_t$. If the government sells B_{t-1} bonds at $t-1$, at t it has to pay bondholders $R_t B_{t-1}$ in the form of principal and interest payments. The right-hand side of Equation 1 lists the various sources of revenue. The government collects taxes worth $N_t \tau_t$ and borrows B goods. The last term on the right-hand side of Equation 1 is seigniorage. M_t represents the total stock of fiat money in the economy at date t. Seigniorage, therefore, is the quantity of goods the government purchases by printing money—that is, $M_t - M_{t-1}$.

The monetary policy in this model economy is identifiable. The central bank controls the money growth rate, at least ostensibly. The nominal money stock evolves over time according to the rule $M_t = \lambda_t M_{t-1}$, where λ is the gross rate of money growth. Using this, seigniorage can be rewritten as

$$v_t M_t \left(1 - \frac{1}{\lambda_t}\right).$$

Thus, unless noted otherwise, we define monetary policy as actions taken by the central bank to change the rate of money growth.

It is helpful to simplify the government's budget constraint by rewriting it in terms of quantities per young person. This involves dividing the expression in Equation 1 by the number of young people each period. Dividing by N_t and using the population growth rule yields

$$(2) \qquad g_t + \frac{R_t}{n_t} b_{t-1} = \tau_t + b_t + v_t m_t \left(1 - \frac{1}{\lambda_t}\right),$$

where $b_t = B_t/N_t$ and $m_t = M_t/N_t$.

Three conditions need to be satisfied for this economy to be in equilibrium at any point t: (1) people consume and acquire money and bonds so as to maximize their lifetime well-being; (2) they take the prices for the consumption good, government bonds, and the value of money as given; and (3) markets clear, in that the supply and demand for consumption goods are equal (likewise for the money and bonds), and the government budget constraint is satisfied.

THE UNPLEASANT MONETARIST ARITHMETIC RESULT

We now describe the simple arithmetic of government budget deficits in the spirit of SW. For convenience, our discussion focuses on the long run, or on steady states, which are equilibrium situations in which government purchases, taxes, bonds, real money balances, and so on (per young person) are invariant with respect to time.[13] In steady states, violation of the government budget constraint at any arbitrary date t results in the present value of government spending, including interest obligations, differing from the present value of government revenues. In a steady state, $b_t = b_{t-1} = b$, and g, τ, R, vm, n, and λ are likewise time-invariant. Equation 2 may be rewritten as

$$(3) \qquad g = \tau + \left(1 - \frac{R}{n}\right) b + vm \left(1 - \frac{1}{\lambda}\right)$$

to form the basis for the SW result.

This article focuses on the case in which the central bank is subservient to the treasury.[14] The government's budget constraint in this steady state is satisfied—that is, Equation 3 holds. A steady-state representation of the unpleasant monetarist arithmetic is as follows: with taxes unchanged, a permanent increase in outstanding government bonds requires a permanent increase in the inflation rate to ensure the government budget constraint is satisfied.

Consider an increase in government purchases of the consumption good, g (holding taxes, τ, and money growth, λ, constant),

funded by an increase in bonds, b, sold to the public. However, the bonds also have interest costs, which in steady state equal $(R/n)b$. The consequences of a bond-financed increase in government spending depend crucially on R/n, the ratio of the real interest rate to the economy's growth rate.[15] The question is whether the costs are less than, equal to, or greater than the revenue the bond sale generates.

First, consider a situation where $R/n > 1$—that is, the real interest rate on government debt exceeds the economy's growth rate. $(R/n)b > b$, or, in plain English, the interest on existing debt exceeds the revenue from the sale of fresh debt, resulting in a revenue shortfall. In equilibrium, the budget balance holds (Equation 3). With τ fixed, the central bank must raise the revenue needed to make up this shortfall. The central bank responds by increasing λ, which raises both seigniorage, $vm\ (1 - 1/\lambda)$, and the inflation rate.

What is the intuition behind this result? With a permanent increase in b, the steady-state comparison indicates that a revenue shortfall will occur. The government must pay for its initial purchases and cover the additional interest expense. Because the economy grows slower than the gross real return on government debt, steady-state interest expenses exceed revenue from the debt issue. Hence, some other revenue source is needed. SW establish that if the necessary condition $R/n > 1$ is satisfied, either higher taxes or more seigniorage is required to cover the increase in government purchases. In the absence of the tax option, unpleasant arithmetic necessarily follows.[16]

To further understand the role played by the $R/n > 1$ stricture, consider the opposite case, in which the government's revenue from the bond sale is large enough to cover the (steady-state) interest expense.[17] Because net interest payments are growing slower than the economy, the government can glean revenue from this bond issue to pay for the government purchases.[18] The excess revenue from the bond sale, defined as $b - (R/n)b$, allows seigniorage requirements and/or taxes to be lowered.

Having established the importance of R/n, we now review the evidence on the relationship between the real interest rate and the growth rate.

HISTORICAL EVIDENCE ON *R/n*

Some argue that the SW result is a theoretical curiosity, that the unpleasant monetarist arithmetic's key prediction is irrelevant because the gross real return on treasury debt is lower than the economy's growth rate. We review the historical evidence on R versus n, since such a comparison is the primary means used to cast doubt on the relevance of the UMA. We then examine the merits of the criticisms.

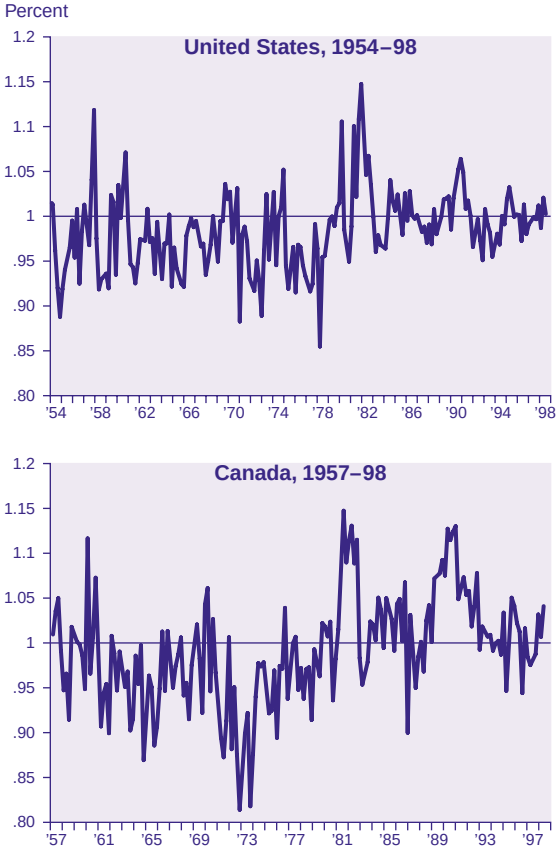
The researcher confronts a number of thorny issues in trying to measure the real interest rate. The principal difficulty is that the rates are generally unobservable. With more countries issuing indexed government debt, some of these measurement issues are mitigated, but the time series on these securities are generally quite short. Hence, real interest rates have to be computed using an observable measure, such as the nominal interest rate, combined with the inflation rate. The question that then arises is whether the GDP deflator or the Consumer Price Index (CPI) is the more appropriate measure of inflation—a question that cannot be conclusively answered. Despite the nettlesome measurement issues, our approach yields the real interest rate on a treasury security held for a specific length of time. In other words, we opt for a measure of the ex post real return paid on a treasury security.

Along the same lines as Champ and Freeman (1994), we plot the ratio R/n for the United States and Canada (*Figure 1*).[19] The ex post real interest rate is measured using the short-term (three month) nominal interest rate and the GDP deflator. We compute the inflation rate for the period in which the short-term government security is outstanding. Figure 1 shows that the real GDP growth rate is usually greater than the real interest rate (the ratio R/n is less than 1). The most notable exceptions occur in Canada from 1988 through 1992, when the real interest rate exceeds the economy's growth rate. Espinosa-Vega and Russell (1998) argue that the appropriate real return is an after-tax real interest rate. According to them, since World War II the after-tax real rate of return on U.S. government debt has been about –0.4 percent, while the average real growth rate of GDP has been about 3.2 percent.

Based on the average gross real return on government debt and the economy's growth rate, one might conclude the unpleasant monetarist arithmetic would have little predictive bite. It is debatable whether the historical evidence from the United States and Canada bears directly on the question Sargent and Wallace raise, which is whether the government can finance a government purchase with a permanent increase in government debt. The United States has never conducted the SW experiment, so drawing inferences about the UMA from the real interest rate and the economy's growth rate

Figure 1

Real Interest Rate and Real GDP Growth



SOURCE: *International Financial Statistics*, various issues.

could be invalid. Suppose, for example, that a permanent increase in government debt were to result in a higher real (after tax) interest rate. In that case, SW's conditions for infeasibility may become likely.

Thus, the historical record on R and n may not be the most damning evidence against the SW result. By this reckoning, the unpleasant monetarist arithmetic is an intellectual curiosity because the experiment is never part of policy, not because its predictive content is invalid. Insofar as the empirical evidence sheds light on the SW predictions, it does so for a setting in which $R > n$ is a necessary condition for obtaining their result. In the next section, we examine whether $R > n$ is in the set of necessary conditions. While the $R > n$ condition may be necessary in the SW setup to generate the UMA, it may not be necessary under slightly more general model specifications.

ADDING STORES OF VALUE

This section examines an economy in which people can hold stores of value other than government bonds and currency. This extension to the basic model economy permits the assessment of the role a restricted asset structure plays in obtaining SW's results. We establish that $R/n > 1$ is a sufficient but not a necessary condition to obtaining the UMA result; unpleasant monetarist arithmetic may result even if $R/n < 1$ holds. Bhattacharya, Guzman, and Smith (1998) develop this analysis more thoroughly.

Consider an economy with a store of value that yields known, fixed units of the consumption good. For concreteness, we refer to this store of value as an investment project. One unit of the consumption good invested in this project yields a units of the consumption good the following period; thereafter, its scrap value is zero.[20] Further assume that investment projects are large, with a minimum size at which they can operate. This minimum is large enough that no individual can finance an investment project.

It seems natural to assume that in such an environment banks would be created to pool people's savings into amounts large enough to fund the investment projects, thereby giving each person an additional store of value. We assume that a bank collects deposits and transforms the goods into an investment project costlessly and that banks operate in a perfectly competitive environment. The bank promises to pay depositors a return one period after the goods are deposited. This return is the same as could be earned by a person with enough resources to invest directly in the project. Since no single person is rich enough to invest directly, we can assume the bank intermediates all investments.

Because banks in this economy are legally required to hold a fraction ($0 < \theta < 1$) of all deposits to meet reserve requirements, the bank's portfolio of assets is divided between money and the investment project. We assume the investment project offers a real return that exceeds the real return of money, that is, $a > n/\lambda$. It is clear the bank will hold an exact fraction θ (never more) of its deposit base in the form of money.

Knowing the returns to these assets (and because providing these banking services costs nothing), it is straightforward to calculate the return to deposits by the bank. The gross real return to currency held by the bank is n/λ, which receives a weight θ in the bank's portfolio. The return to the investment project is a, with corresponding weight $(1 - \theta)$. Hence, the return to deposits is a weighted sum of the returns to each of the bank's asset holdings:

(4) $$\theta\frac{n}{\lambda} + (1-\theta)a.$$

We assume that people can continue to purchase riskless government debt directly and that there are no reserve requirements on debt holdings. A person's savings will be divided in two: part will be used to buy government bonds (b), and the remainder will be deposited with the bank (d)—that is, $s = b + d$. For people to hold government debt willingly, its return must be at least as great as the return to deposits. Likewise, people will hold deposits if their return is at least as great as the return to government debt. If both government debt and bank deposits are to be held,

(5) $$R = \theta\frac{n}{\lambda} + (1-\theta)a$$

must result. Equation 5 is what is often called a no-arbitrage condition. More generally, multiple riskless assets will be held only if they bear identical rates of return.[21]

We proceed in two steps. The first step requires that the returns to government bonds, the investment project, and money be ranked. If $a > n > n/\lambda$ holds, $R > n/\lambda$ must also hold.[22] To verify this, suppose $a > n > n/\lambda$, but $R < n/\lambda$. It follows from Equation 5 that $\theta(n/\lambda) + (1-\theta)a < n/\lambda$ must also hold. This expression reduces to $a < n/\lambda$, which contradicts our original assumption that $a > n/\lambda$. Hence, $a > n/\lambda$ implies $R > n/\lambda$. If the investment project pays a higher real return than money, the real return to government debt must exceed the return to fiat money. Thus, government debt is more expensive (to the government) than money as a means of deficit financing. Treasury debt requires an explicit (nominal) interest payment while money balances do not. (Note that $a > n > n/\lambda$ does not imply $R > n$.)

Our second step examines the case in which the government permanently increases the ratio of government bonds to cash balances. In this instance, the government is selling government debt and buying money—an open market sale. To determine what this means for the government's unified budget constraint, we return to Equation 3, setting $\tau = 0$ to get

$$g = \left(1 - \frac{R}{n}\right)b + vm\left(1 - \frac{1}{\lambda}\right).$$

With $a > n > n/\lambda$, we know $R/n > n/\lambda$ or $(1 - R/n) > (1 - 1/\lambda)$. We have completed our task if we can show that the UMA holds even if $R/n < 1$ holds. Suppose $R/n < 1$, and recall that $s = d + b$. If the government uses an open market sale (denoted Δb) to increase b, and if these bonds are held by the public, d must fall for given s. Since a fraction θ of d constitutes real money demand, we know vm (the seigniorage tax base) must fall by Δvm. So an increase in b raises revenue by the amount $(1 - R/n)\Delta b$ (recall that we are assuming $R/n < 1$). However, for fixed λ, the revenue from money creation, $\Delta vm(1 - 1/\lambda)$, falls. Bhattacharya, Guzman, and Smith (BGS) prove that when $(1 - R/n) < (1 - 1/\lambda)$, the loss in seigniorage exceeds the revenue from bonds.[23] This revenue shortfall has to be made up somehow. One possibility is to raise the money growth, λ, which would, of course, raise the inflation rate. Here is what is happening: the government is raising revenue from the sale of bonds but is losing seigniorage (because the bond sale crowds out money holdings and reduces the inflation tax base). With money growth constant, it is possible the net effect is that the bond sale reduces overall steady-state revenues.

Thus, the two steps establish that a permanent increase in government debt cannot be financed by a permanent increase in government bonds. The first step establishes that the real return on government bonds exceeds the real return on fiat money. With $R > n/\lambda$ and $a > n$, BGS show these two results are sufficient for the UMA. In plain English, the unpleasant monetarist arithmetic could hold even if the real return to government bonds is lower than the economy's growth rate. The key proviso is that there exists a store of value with a real return higher than the economy's growth rate. The open market sale means the government must cover the net interest expenses of the larger stock of debt while reducing the quantity of real money balances. Higher inflation is necessary to pay for these expenses, potentially even when the real return on the debt is lower than the economy's growth rate.

The BGS finding is important because it means the UMA result can be obtained with a set of necessary conditions that does not include $R/n > 1$ (condition 2). Indeed, in the BGS framework, the UMA is a possibility as long as there is an intermediated asset with a real return exceeding the economy's growth rate. Since this last condition does not require that R/n be greater than 1, BGS may have eliminated its need. This undercuts the criticisms leveled by Darby (1984) and others against the "unrealistic" $R/n > 1$ condition.

In the BGS economy, the gross real return to the investment project, a, is constant. Bhattacharya and Kudoh (1998) consider a neoclassical production economy in which the in-

A Seigniorage Laffer Curve

The Laffer curve was originally developed to show that income tax rates can get so high people start to choose nonmarket activities over working. If the rate at which people drop out of work is fast enough relative to the rate at which income taxes are raised, income tax revenue—the product of people's income (the tax base) and the tax rate on it—could decline.

Monetary economists have used similar reasoning to explore the effect increasing the inflation tax has on seigniorage. Recall, the revenue earned from money creation is

$$vm\left(1-\frac{1}{\lambda}\right);$$

vm is the seigniorage tax base, and $1 - 1/\lambda$ is the tax rate.

First, consider the relationship between changes in the money growth rate and the level of real seigniorage. The idea is that faster money growth is associated with higher inflation rates. Holding everything else (especially the seigniorage tax base) constant, higher inflation would produce higher real seigniorage. However, the inflation rate is inversely related to the gross real return to holding money. If people hold less money in response to the higher inflation, it is clear the tax base is declining in the face of a higher tax rate. Inflation rates can get so high that people start to eschew money. In short, vm declines.

The Laffer curve description fits the following scenario. When the inflation tax rate is low, the decline in the seigniorage tax base is small and the product (real seigniorage) rises with increases in the inflation rate. However, it is possible that at high money growth rates (and hence at sufficiently high inflation rates), people will reduce their money holdings so much that real seigniorage may fall with further increases in the inflation rate. When money growth rates and real seigniorage are positively related, they are on the good side of the Laffer curve. Conversely, when a decrease in real seigniorage accompanies an increase in the money growth rate, they are on the bad side of the Laffer curve.

Second, consider the relationship between reserve requirements and real seigniorage. Again, holding everything else constant, an increase in the reserve requirement will raise the seigniorage tax base, resulting in greater seigniorage. However, if the rate of return to other stores of value exceeds that of money, an increase in the reserve requirement drives a greater wedge between the return to money relative to other stores of value. If reserve requirements rise enough, people have an incentive to move their savings from banks to stores of value that do not face reserve requirements. This action could result in a smaller seigniorage tax base. As with the money growth rate, therefore, the relationship between the reserve requirement ratio and real seigniorage may be inverse-U shaped.

vestment project exhibits diminishing marginal returns.[24] As the government increases its reliance on bonds, investment in the project gets crowded out, raising the return on the project and thereby raising the return to bonds, R. Because of the no-arbitrage condition, government bond financing becomes costlier as interest expenses rise with R. Bhattacharya and Kudoh find that even when the real return to capital (analogous to a) is lower than the economy's growth rate, the UMA is still a possibility. As such, they show that the SW/BGS results extend to a more general economic model.

DIFFERENT MONETARY POLICY TOOLS

Sargent and Wallace restrict the central bank to a single policy tool, the money growth rate. However, real-world central banks have other means of raising seigniorage. For example, many directly control the reserve requirement ratio. This gives rise to a broader question: faced with an increase in treasury debt, is it better for the central bank to change the reserve ratio or change the money growth rate? For the purpose of the UMA, does it matter how monetary policy is implemented? Is it possible that the UMA may not result if the central bank changes the reserve ratio (instead of the money growth rate) to raise the required seigniorage?

Freeman (1987) identifies the optimal way for the central bank to raise seigniorage. Abstracting from government debt, he shows that monetary policy could mimic a lump-sum tax. Set the reserve requirement equal to the ratio of government purchases to output. By confiscating this amount of real money balances, the government could fund its purchases. Confiscation would be achieved by making these money balances worthless—that is, by letting money grow at an infinite rate.[25] Thus, Freeman shows that a combination policy using both money growth rates and the reserve requirement ratio would be best.

Bhattacharya and Haslag (1999) study a production economy, similar to the Bhattacharya and Kudoh (1998) economy, in which the central bank controls two monetary policy tools. The central bank applies either the reserve requirement or the money growth rate, holding the other constant, to affect the level of seigniorage. By changing the reserve requirement, the central bank alters the seigniorage tax base, while keeping the seigniorage tax rate constant. Bhattacharya–Haslag quantify the change in the steady-state level of real seigniorage following a change of central bank tool.

People store for future consumption by holding money, government bonds, and investment projects. The central bank could raise the seigniorage needed to cover a permanent increase in government bonds by raising the money growth rate or by lowering the reserve requirement ratio. Because a person's lifetime consumption increases as reserve requirements are lowered, the reserve requirement policy is a pleasant monetarist arithmetic.

In addition to addressing the possibility that monetary policy may be implemented in multiple ways, the Bhattacharya–Haslag exercise underscores the importance of the seigniorage Laffer curve in this discussion. SW's unpleasant monetarist arithmetic occurs because the economy is on the "good"' side of the Laffer curve with respect to the money growth rate: an increase in the money growth rate generates an increase in seigniorage. (The box entitled "A Seigniorage Laffer Curve" discusses the Laffer curve as it applies to monetary policy.) In contrast, the Bhattacharya–Haslag results indicate the econ-

omy is on the "bad" side of the Laffer curve with respect to the reserve requirement: a decrease in the reserve requirement raises seigniorage.

For the Bhattacharya–Haslag result, the intuition is straightforward. With higher reserve requirements, fewer of the deposited goods are put into the investment project.[26] As such, investment projects are crowded out of the bank's portfolio. Diminishing marginal returns play a crucial role here. The gross real return on the investment project would rise. If government bonds are held, their return would rise, too. It follows that the government's interest expense would rise: there are more government bonds, and the interest rate on each bond is higher. The increase in the reserve requirements drives up interest expenses, which will require more seigniorage, which will, in turn, require a further increase in the reserve requirement, and so on. Bhattacharya–Haslag find that lowering reserve requirements increases investment in the project enough that the gross real return on government debt declines, permitting the smaller tax base to finance the increase in government expenses. Thus, their findings suggest that how monetary policy is implemented does impact the unpleasant monetarist arithmetic, in the sense that people prefer lower reserve requirements to faster money growth when the central bank must pay for a permanent increase in government bonds. Put differently, a more "pleasant" monetarist arithmetic may be observed if the central bank reduces the reserve ratio to raise the required revenue.

CONCLUDING REMARKS

This article reviews some recent developments in the unpleasant monetarist arithmetic literature, with a focus on the studies that adopt the Sargent–Wallace approach of making the central bank subservient to the treasury. We ignore the literature that explores the game of chicken between these two entities.

This survey highlights two main developments in the literature. The SW result seems to rest squarely on the proviso that the real interest rate on government debt is greater than the economy's growth rate. However, the data appear to relegate this result to that of a theoretical curiosity; for most of the postwar period, the real interest rate has been below the growth rate in both the United States and Canada. One recent development in the literature shows that for the SW result to hold, it is not necessary that the real return on government debt exceed the economy's growth rate. If there is an asset with a real return that exceeds the economy's growth rate, and if government debt offers a positive nominal interest rate, the SW result is possible even if the real return on government debt is less than the economy's growth rate. Both these sufficient conditions seem empirically plausible; for many countries, the average real return on equities is, on average, above the economy's growth rate. Thus, a subservient central bank could still be required to support the treasury's financing needs, even if the real return on government debt is quite low.

Second, we examine a case in which the central bank controls more than one policy tool, to determine whether the unpleasant monetarist arithmetic depends on which tool is used. Numerical analyses indicate that a permanent increase in government debt requires faster money growth, at least for low to medium money growth rates, but lower reserve requirements. The findings are consistent with the notion that reserve requirements are a blunt instrument. Movements in the reserve requirement ratio directly crowd out capital from a bank's portfolio. Movements in the money growth rate do not have such a direct impact on the distribution of the bank's assets. Because the rates of return on these assets are affected, our analysis suggests that people would prefer lower reserve requirements, and thus lower real rates of return, to faster money growth to finance the government's bond issue.

A caveat is necessary. In this article, we provide a status report on the unpleasant monetarist arithmetic, under the explicit assumption the central bank is completely subservient to the treasury. This leaves unanswered the question of to what degree (if any) the central bank should be subservient to the treasury. Answering such a question would require delving into the myriad strategic considerations that determine the degree to which the central bank is independent of the treasury. Doubtless, this is interesting material for future work.[27]

NOTES

The authors wish to thank Helle Bunzel, Tim Fuerst, Noritaka Kudoh, and, especially, Evan Koenig, Mark Wynne, and Carlos Zarazaga for helpful comments on an earlier version of this article.

1 Of course, well-defined normative criteria (objective functions) are needed for judging efficacy. The appropriateness of government objective function(s) is outside the scope of this article.

2 The basic ideas had already been presented by Metzler (1951), Patinkin (1965), and especially Christ (1968).

[3] A standard game of chicken involves the following. Two players, each in a car, face each other, separated by a distance of, say, 100 yards. Someone blows a whistle and the cars start rushing toward each other. If neither player moves out of the other's way, both would die. So one player has to give. The one that gives way to the other is the "chicken"; the other player wins. Sargent (1987, 176) attributes the idea of viewing monetary–fiscal policy interactions as a game of chicken to Wallace.

[4] Thinking about these issues as a game of chicken is enlightening in regard to Europe's recent move toward a single currency. The issue of deficit financing arises because the treasuries in the eleven countries would appear to lose the game of chicken against the European central bank.

[5] This policy experiment was popular among researchers embroiled in the monetarists vs. Keynesians debate.

[6] There is no reason, a priori, to believe that an independent central bank would raise seigniorage to meet the treasury's interest obligations. An alternative is for the treasury to raise taxes. A good analogy is the example of the Federal Reserve and the state of Texas. There is no evidence the Fed creates money to meet the state's obligations, although the state is clearly a passive beneficiary of unexpected increases in money growth rates brought about by the Fed.

[7] In this article, we stick to the version of unpleasant monetarist arithmetic Sargent (1987) presents. In particular, we restrict our focus to a comparison of steady states (see the section titled "The Unpleasant Monetarist Arithmetic Result"), which differs slightly from SW (1981). SW compare the inflation rate for two cases: one in which the government finances its deficit with money creation today and one in which the government issues bonds to finance the deficit and is eventually forced to monetize the deficit. The inflation rates differ in the two cases; indeed, SW derive conditions in which the inflation rate is higher in the bond-finance case.

We follow the approach adopted in Sargent (1987). There, the comparisons are conducted on stationary equilibria. Sargent writes: "The higher the stationary value of interest-bearing government debt b, the lower the rate of return on currency, that is, the higher the inflation rate. This is the foundation of Sargent and Wallace's result" (147). As such, we derive conditions in which higher inflation is part of the policy package accompanying an increase in government bonds.

[8] To see this, let population growth in this economy be represented by $N_t = nN_{t-1}$, so that the gross population growth rate is n. That is, if n were equal to 1 at date $t - 1$, it would mean the adult population replaced itself one for one with children that period or the population did not grow between dates $t - 1$ and t. Aggregate income (GDP) is the product of the number of young people and their endowments; that is, $N_t y$. Thus, the aggregate income growth rate is given by $N_t y/N_{t-1} y$. With a constant endowment per young person, y, the income growth rate reduces to n, the same as the population growth rate.

[9] We have eliminated the possibility of intergenerational loans. The old would never loan anything to the young because repayment would occur after the old have died. (There are no dynasties that could enforce repayment.) The young would never loan anything to the old because there is no way to enforce contracts with a dead person. See Samuelson (1958) for details.

[10] Put differently, v_t is the inverse of the price level at date t in this economy.

[11] If the choice is between holding interest-bearing bonds or holding non-interest-bearing money, no person will hold money unless its real rate of return is the same as the real return on bonds. In other words, we need to specify the reason people hold money even when better stores of value are available. We return to these issues in the section "Adding Stores of Value."

[12] Fiat money is nothing but intrinsically worthless pieces of paper money that are inconvertible—that is, the government does not promise to convert the money into the consumption good. See Wallace (1980).

[13] Since everything in the economy (except the price level) in a steady state is time-invariant, the subscript t loses any meaning and is therefore dropped.

[14] Central bank independence can be defined in terms of the game of chicken between the treasury and the central bank described above. Suppose the fiscal authority chooses it policies first (τ and b), independent of the central bank. The central bank, having lost the game of chicken, sets λ to ensure that current and future money creation pays for all the treasury's future interest obligations and the government's expenditures. This is our definition of a subservient central bank. Aiyagari and Gertler (1985) label this a non-Ricardian regime. See also Canzoneri, Cumby, and Diba (1998) for alternative classifications in terms of fiscal-dominant and money-dominant regimes. By their definition, the postwar United States has a money-dominant regime. It is important to note that a fiscal-dominant regime in the sense of Canzoneri, Cumby, and Diba (1998) is not the same as a fiscal leadership regime (a term SW use), which in turn differs from a non-Ricardian regime.

[15] Abel (1992) also discusses the consequences of government financing when the real return on government debt exceeds the economy's growth rate. Abel focuses on deficit financing's impact on the capital stock.

[16] Sargent and Wallace go one step further, to show that if people are forward-looking, they will know future money creation is necessary to fund the increase in the government's debt; as a result, current inflation will rise.

[17] Sargent and Wallace (1981), Darby (1984), and Miller and Sargent (1984) also recognize this case and discuss both sides of the debt-financing issue.

[18] Describing the transition from one steady state to the other can be thought of as a case in which the ratio of

public debt to GDP is declining over time. With a declining ratio of debt to aggregate income, principal and interest payments associated with the bond issue become a smaller fraction of the economy. It follows that the growing economy can absorb the debt obligations without having to rely on additional taxes or seigniorage to pay for the initial purchase.

[19] We focus on the United States and Canada because it is the data for these countries that have been used to argue against the SW result. We also have data for Russia and Brazil, although there are only four years of data from Russia and there is a five-year gap in Brazil's R/n ratio. The data are available from the authors upon request.

[20] A real-world analog of this would be a time deposit or certificate of deposit.

[21] It is important to note that the reserve requirement on money holdings is singularly responsible for creating a wedge between the return to government debt and the return to the investment project.

[22] This combines two previously discussed stipulations: investment projects must yield returns that are higher than both the economy's growth rate and the return on money.

[23] The important point here is that this is true even though in this regime $R/n < 1$ (condition 3) is not satisfied. This is also the sense in which printing money is the cheaper option.

[24] Taking this to its natural limit, if the economy could put an infinite quantity of goods into the investment project, the return on the last unit would be zero. Conversely, if the quantity of goods in the investment project is close to its minimum-size requirement, an additional unit of the good put into the project will offer returns that greatly exceed the economy's growth rate.

[25] A technical consideration arises when the money growth rate is set equal to infinity. The Freeman policy prescription works if the money growth rate is some very large, finite number so that the value of money is close to zero.

[26] Recall that people save the same amount regardless of the rate of return. Thus, total saving is taken as given.

[27] In an interesting paper, Carlstrom and Fuerst (forthcoming) examine the rules in the game of chicken in terms of how changes in timing affect the relationship between the central bank and the treasury.

REFERENCES

Aiyagari, S. Rao, and Mark Gertler (1985), "The Backing of Government Bonds and Monetarism," *Journal of Monetary Economics* 16 (July): 19–44.

Abel, Andrew B. (1992), "Can the Government Roll Over Its Debt Forever?" Federal Reserve Bank of Philadelphia *Business Review*, November/December, 3–18.

Bhattacharya, Joydeep, Mark G. Guzman, and Bruce D. Smith (1998), "Some Even More Unpleasant Monetarist Arithmetic," *Canadian Journal of Economics* 31 (August): 596–623.

Bhattacharya, Joydeep, and Joseph H. Haslag (1999), "Seigniorage in a Neoclassical Economy: Some Computational Results," Federal Reserve Bank of Dallas Research Working Paper no. 99-01 (Dallas, January).

Bhattacharya, Joydeep, and Noritaka Kudoh (1998), "Tight Money Policies and Inflation Revisited" (Unpublished manuscript, State University of New York, Buffalo).

Canzoneri, Matthew B., Robert E. Cumby, and Behzad T. Diba (1998), "Is the Price Level Determined by the Needs of Fiscal Solvency?'' NBER Working Paper Series, no. 6471 (Cambridge, Mass.: National Bureau of Economic Research, March).

Carlstom, Charles, and Timothy Fuerst (forthcoming), "The Fiscal Theory of the Price Level," Federal Reserve Bank of Cleveland *Economic Review*.

Champ, Bruce, and Scott Freeman (1994), *Modeling Monetary Economies* (Boston: John Wiley & Sons).

Christ, Carl F. (1968), "A Simple Macroeconomic Model with a Government Budget Restraint," *Journal of Political Economy* 76 (January/February): 53–67.

Darby, Michael (1984), "Some Pleasant Monetarist Arithmetic," Federal Reserve Bank of Minneapolis *Quarterly Review*, Spring, 15–20.

Espinosa-Vega, Marco A., and Steven Russell (1998), "Can Higher Inflation Reduce Real Interest Rates in the Long Run?" *Canadian Journal of Economics* 31 (February), 92–103.

Freeman, Scott (1987), "Reserve Requirements and Optimal Seigniorage," *Journal of Monetary Economics* 19 (March), 307–14.

Metzler, L. (1951), "Wealth, Saving, and the Rate of Interest," *Journal of Political Economy* 59 (April): 93–116.

Miller, Preston, and Thomas Sargent (1984), "A Reply to Darby," Federal Reserve Bank of Minneapolis *Quarterly Review,* Spring, 21–26.

Patinkin, Don (1965), *Money, Interest, and Prices*: *An Integration of Monetary and Value Theory*, 2nd ed., (New York: Harper & Row).

Samuelson, Paul (1958), "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance

of Money," *Journal of Political Economy* 66 (December): 467–82.

Sargent, Thomas (1987), *"Dynamic Macroeconomic Theory"* (Cambridge: Harvard University Press).

Sargent, Thomas J., and Neil Wallace (1981), "Some Unpleasant Monetarist Arithmetic," Federal Reserve Bank of Minneapolis *Quarterly Review*, Fall, 1–17.

Wallace, Neil (1980), "The Overlapping Generations Model of Fiat Money," in *Models of Monetary Economies*, ed. J. Karaken and N. Wallace (Minneapolis: Federal Reserve Bank of Minneapolis), 49–82.